PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-202973

(43)Date of publication of application : 19.07.2002

(51)Int.Cl. G06F 17/30

(21)Application number : 2001-291628

(22)Date of filing : 25.09.2001

(71)Applicant : MATSUSHITA ELECTRIC IND CO LTD

(72)Inventor : SHIMOJIMA TAKASHI
ITO MASAO
TSURUBAYASHI TAKESHI
KATAYAMA OSAMU
NAKAI SHINICHI

(30)Priority

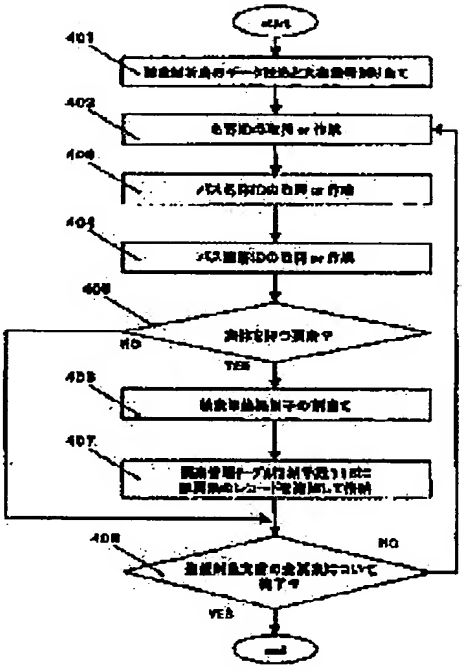
Priority number : 2000325286 Priority date : 25.10.2000 Priority country : JP

(54) STRUCTURED DOCUMENT MANAGEMENT DEVICE

(57)Abstract:

PROBLEM TO BE SOLVED: To provide a structured document management device capable of performing various retrievals designated in logical structure.

SOLUTION: In a document management system for handling structured documents, the information for specifying a logical structure position is managed with a pass name with tag names continuously described in order from the top hierarchy and a pass hierarchy with appearing orders of each hierarchy of the pass name continuously described, whereby various structured document retrievals can be realized.



LEGAL STATUS

[Date of request for examination] 05.10.2001

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2002－202973
（P2002－202973A）
(43)公開日　平成14年７月19日(2002.7.19)

(51)Int.Cl.'	識別記号	ＦＩ		ﾃｰﾏｺｰﾄﾞ(参考)
Ｇ０６Ｆ　17/30	１４０	Ｇ０６Ｆ　17/30	１４０	５Ｂ０７５
	４１９		４１９Ａ	

審査請求　有　　請求項の数26　ＯＬ　（全 43 頁）

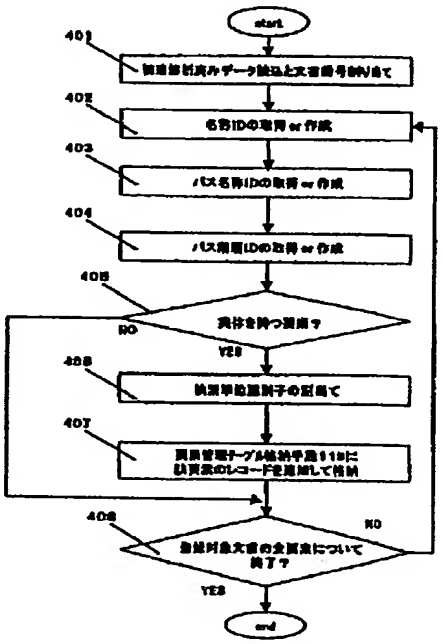
(21)出願番号　特願2001－291628(P2001－291628)
(22)出願日　平成13年９月25日(2001.9.25)
(31)優先権主張番号　特願2000－325286(P2000－325286)
(32)優先日　平成12年10月25日(2000.10.25)
(33)優先権主張国　日本（ＪＰ）

(71)出願人　000005821
松下電器産業株式会社
大阪府門真市大字門真1006番地
(72)発明者　下島　崇
大阪府門真市大字門真1006番地　松下電器産業株式会社内
(72)発明者　伊藤　正雄
大阪府門真市大字門真1006番地　松下電器産業株式会社内
(74)代理人　100097445
弁理士　岩橋　文雄　（外２名）

最終頁に続く

(54)【発明の名称】　構造化文書管理装置

(57)【要約】
【課題】　様々な論理構造を指定した検索をすることのできる構造化文書装置を提供する。
【解決手段】　構造化文書を扱う文書管理システムにおいて、論理構造位置を特定するための情報を、最上位階層から順にタグ名を連ねて記述したパス名称と、パス名称の各階層の出現順序を連ねて記述したパス階層で管理することにより、様々な構造化文書検索を実現することができる。



【特許請求の範囲】

【請求項１】　構造化文書を扱う文書管理装置において、構造化文書の入力を行う構造化文書入力手段と、前記構造化文書入力手段により取り込んだ構造化文書を解析し該構造化文書の木構造を生成する構造解析手段と、前記構造解析手段により木構造で表現された構造化文書において、各要素実体を識別する検索単位識別子と、各要素実体の前記木構造における位置を表現した要素実体位置識別子と、前記検索単位識別子から前記要素実体位置識別子を特定するために、少なくとも前記検索単位識別子と関係する前記要素実体位置識別子を対応付けた要素管理テーブルを作成する構造情報作成手段と、文字列検索を行うための文字列索引を作成する文字列索引作成手段と、検索条件を入力する検索条件入力手段と、前記検索条件入力手段で入力された検索条件に該当する前記要素実体位置識別子を特定する検索条件解析手段と、前記文字列索引作成手段で作成された文字列索引を用いて検索条件に該当する文字列を有する各要素実体の前記検索単位識別子を特定する文字列索引検索手段と、前記文字列索引検索手段で特定した検索単位識別子を基に前記要素管理テーブルを参照して対応する要素実体位置識別子を求め、前記要素実体位置識別子と前記検索条件解析手段により求めた前記要素実体位置識別子とが一致する検索単位識別子のみを抽出する構造照合手段を備えた構造化文書管理装置。

【請求項２】　構造化文書を扱う文書管理装置において、構造化文書の入力を行う構造化文書入力手段と、前記構造化文書入力手段により取り込んだ構造化文書を解析し該構造化文書の木構造を生成する構造解析手段と、前記構造解析手段により木構造で表現された構造化文書において、各要素実体を識別する検索単位識別子と、各要素実体に至るタグ名を階層順に連ねたパス名称を識別するパス名称ＩＤと、同一の親ノードを持ち同一な名称を持つタグの同一階層内での出現順序を階層順に連ねたパス階層を識別するパス階層ＩＤと、前記検索単位識別子から前記パス名称ＩＤと前記パス階層ＩＤを特定するために、少なくとも前記検索単位識別子と関係する前記パス名称ＩＤ及びパス階層ＩＤを対応付けた要素管理テーブルを作成する構造情報作成手段と、文字列検索を行うための文字列索引を作成する文字列索引作成手段と、検索条件の入力を行う検索条件入力手段と、前記検索条件入力手段で入力された検索条件に該当する前記パス名称ＩＤ、前記パス階層ＩＤの少なくともいずれか１つを特定する検索条件解析手段と、前記文字列索引作成手段で作成された文字列索引を用いて検索条件に該当する文字列を有する各要素実体の前記検索単位識別子を特定する文字列索引検索手段と、前記文字列索引検索手段で特定した検索単位識別子を基に前記要素管理テーブルを参照して対応するパス名称ＩＤ又はパス階層ＩＤを求め、前記パス名称ＩＤ又は前記パス階層ＩＤと前記検索条件解析手段により求めたパス名称ＩＤ又はパス階層ＩＤとが一致する検索単位識別子のみを抽出する構造照合手段を備えた構造化文書管理装置。

【請求項３】　構造化文書を扱う構造化文書管理装置において、構造化文書の入力を行う構造化文書入力手段と、前記構造化文書入力手段により取り込んだ構造化文書を解析し該構造化文書の木構造を生成する構造解析手段と、前記構造解析手段により生成された木構造からタグ名を識別する名称ＩＤと、各要素実体を識別する検索単位識別子と、前記検索単位識別子から前記名称ＩＤを特定するために、少なくとも前記検索単位識別子と関係する前記名称ＩＤを対応付けた要素管理テーブルを作成する構造情報作成手段と、文字列検索を行うための文字列索引を作成する文字列索引作成手段と、検索条件を入力する検索条件入力手段と、前記検索条件入力手段で入力された検索条件に該当する前記名称ＩＤを特定する検索条件解析手段と、前記文字列索引作成手段で作成された文字列索引を用いて検索条件に該当する文字列を有する各要素実体の前記検索単位識別子を特定する文字列索引検索手段と、前記文字列索引検索手段で特定した検索単位識別子を基に前記要素管理テーブルを参照して対応する名称ＩＤを求め、前記名称ＩＤと前記検索条件解析手段で求めた名称ＩＤとが一致する検索単位識別子のみを抽出する構造照合手段を備えた構造化文書管理装置。

【請求項４】　文字列検索結果一覧や各要素実体表示のためのデータを作成する結果作成手段と、前記結果作成手段で作成された検索結果を端末に表示する結果表示手段とを有することを特徴とする請求項１から３に記載の構造化文書管理装置。

【請求項５】　構造化文書の入力を行う構造化文書入力手段と、前記構造化文書入力手段により取り込んだ構造化文書を解析し該構造化文書の木構造を生成する構造解析手段と、前記構造解析手段により木構造で表現された構造化文書において、各要素実体を識別する検索単位識別子と、各要素実体に至るタグ名を階層順に連ねたパス名称を識別するパス名称ＩＤと、同一の親ノードを持ち同一な名称を持つタグの同一階層内での出現順序を階層順に連ねたパス階層を識別するパス階層ＩＤと、前記検索単位識別子から前記パス名称ＩＤ及び前記パス階層ＩＤを特定するために、少なくとも前記検索単位識別子と関係する前記パス名称ＩＤ及びパス階層ＩＤを対応付けた要素管理テーブルを作成する構造情報作成手段とを備えた構造化文書登録装置。

【請求項６】　構造化文書の木構造が変化した場合に、要素管理テーブルに記録されたパス名称ＩＤ、パス階層ＩＤのうち、変更が必要なＩＤを更新することを特徴とする請求項２記載の構造化文書管理装置。

【請求項７】　構造化文書の木構造が変化した場合に、要素管理テーブルに記録されたパス名称ＩＤ、パス階層ＩＤのうち、変更が必要なＩＤを更新することを特徴と

する請求項５記載の構造化文書登録装置。

【請求項８】　構造化文書の入力を行う構造化文書入力手段と、前記構造化文書入力手段により取り込んだ構造化文書を解析し該構造化文書の木構造を生成する構造解析手段と、前記構造解析手段により生成された木構造からタグ名を識別する名称ＩＤと、各要素実体を識別する検索単位識別子と、前記検索単位識別子から前記名称ＩＤを特定するために、少なくとも前記検索単位識別子と関係する前記名称ＩＤを対応付けた要素管理テーブルを作成する構造情報作成手段とを備えた構造化文書登録装置。

【請求項９】　各要素実体内部にさらにタグに囲まれた要素実体（子要素）を含む構造化文書の索引作成において、各要素実体から所定の文字数で取り出した文字列が前記タグにまたがる場合は、該子要素を識別する独自の検索単位識別子を取得し、該文字列と該文字列の各文字の属する要素実体を識別する検索単位識別子と前記タグを取り除いた要素実体内での該文字列の位置を示す文字位置識別子とから成る検索用文字列索引を生成することを特徴とする文字列索引作成装置。

【請求項１０】　予め数値であることを定義しているタグに囲まれた文字列を含む構造化文書の索引作成において、該タグに囲まれた文字列を識別する独自の検索単位識別子を取得し、該タグに囲まれた文字列を数値データに変換し、前記検索単位識別子と前記数値データとを対応付けた数値型索引を作成する数値型索引作成手段を有していることを特徴とする請求項９記載の文字列索引作成装置。

【請求項１１】　所定の条件に該当する文字列を検索する場合において、タグ名を識別する名称ＩＤと、各要素実体に至るタグ名を階層順に連ねたパス名称を識別するパス名称ＩＤと、同一の親ノードを持ち同一な名称を持つタグの同一階層内での出現順序を階層順に連ねたパス階層を識別するパス階層ＩＤと、各要素実体を識別する検索単位識別子と、前記検索単位識別子から前記名称ＩＤを特定するために、少なくとも前記検索単位識別子と関係する前記名称ＩＤを対応付けた要素管理テーブルまたは、前記検索単位識別子から前記パス名称ＩＤと前記パス階層ＩＤを特定するために、少なくとも前記検索単位識別子と関係する前記パス名称ＩＤ及びパス階層ＩＤを対応付けた要素管理テーブルの少なくともいずれか一方を記憶するデータ格納部と、検索条件の入力を行う検索条件入力手段と、前記検索条件入力手段で入力された検索条件から検索条件に該当する前記名称ＩＤ、前記パス名称ＩＤ、前記パス階層ＩＤの少なくともいずれか１つ（ＩＤ１）を特定する検索条件解析手段と、検索条件に該当する文字列を有する前記検索単位識別子を求める文字列索引検索手段と、前記文字列索引検索手段で特定した検索単位識別子を基に前記要素管理テーブルを参照して対応する名称ＩＤ、パス名称ＩＤ、パス階層ＩＤの少なくともいずれか１つ（ＩＤ２）を求め、前記ＩＤ２と前記検索条件解析手段により求めたＩＤ１とが一致する検索単位識別子のみを抽出する構造照合手段を備えた文字列検索装置。

【請求項１２】　予め数値であることを定義しているタグに囲まれた文字列を含む構造化文書の数値範囲検索において、前記タグに囲まれた文字列を識別する独自の検索単位識別子と前記タグに囲まれた文字列を数値に変換した数値データとを対応付けた数値型索引を参照し、検索条件に該当する前記検索単位識別子を抽出する数値型索引検索手段を有することを特徴とする請求項１１記載の文字列検索装置。

【請求項１３】　木構造で表現した構造化文書の登録方法において、構造化文書を読み込むステップと、各要素実体に至るタグ名を階層順に連ねたパス名称を識別するパス名称ＩＤと、同一の親ノードを持ち同一な名称を持つタグの同一階層内での出現順序を階層順に連ねたパス階層を識別するパス階層ＩＤを取得するステップと、要素実体を有するか否かを判断するステップと、各要素実体を識別する検索単位識別子を取得するステップと、前記検索単位識別子から前記パス名称ＩＤ及び前記パス階層ＩＤを特定するために、少なくとも前記検索単位識別子と関係する前記パス名称ＩＤ及びパス階層ＩＤを対応付けた要素管理テーブルを作成するステップを有するプログラムを記録した可搬型媒体。

【請求項１４】　木構造で表現した構造化文書の登録方法において、構造化文書を読み込むステップと、タグ名を識別する名称ＩＤを取得するステップと、要素実体を有するか否かを判断するステップと、各要素実体を識別する検索単位識別子を取得するステップと、前記検索単位識別子から前記名称ＩＤを特定するために、少なくとも前記検索単位識別子と関係する前記名称ＩＤを対応付けた要素管理テーブルを作成するステップを有するプログラムを記録した可搬型媒体。

【請求項１５】　要素実体内部にさらにタグに囲まれた要素実体（子要素）を有する構造化文書の文字索引の生成方法について、構造解析済みデータを読み込むステップと、要素実体を有するか否かをチェックするステップと、要素実体を識別するための検索単位識別子を取得するステップと、前記子要素を含むか否かを調べるステップと、該子要素を識別する検索単位識別子を取得するステップと、要素実体から１以上の所定文字数を単位とする文字列を取り出すステップと、前記文字列の各文字の属する検索単位識別子を求めるステップと、該文字列及び該文字列の各文字の属する前記検索単位識別子及び前記タグを取り除いた要素実体内での当該文字列の位置を示す文字位置識別子を有する検索文字列索引を生成するステップとを有するプログラムを記録した可搬型媒体。

【請求項１６】　構造化文書の数値検索用索引生成方法について、構造解析済みデータを読み込むステップと、

予め数値であることを定義しているタグに囲まれた文字列であるか否かを判断するステップと、数値であることを定義したタグに囲まれた文字列を識別するための検索単位識別子を取得するステップと、該文字列を数値に変換するステップと、前記検索単位識別子と前記変換された数値とからなる数値型索引を生成するステップを有するプログラムを記録した可搬型媒体。

【請求項１７】　構造化文書の検索方法について、検索条件を読み込むステップと、前記検索条件に該当するタグ名を識別する名称ＩＤ又は、各要素実体に至るタグ名を階層順に連ねたパス名称を識別するパス名称ＩＤ又は、同一の親ノードを持ち同一な名称を持つタグの同一階層内での出現順序を階層順に連ねたパス階層を識別するパス階層ＩＤの少なくともいずれか１つのＩＤ（以下、ＩＤ１）に変換するステップと、検索条件に該当する文字列を有する各要素実体を識別する検索単位識別子（以下、ＩＤ２）を特定するステップと、前記ＩＤ２から前記名称ＩＤ、前記パス名称ＩＤ、前記パス階層ＩＤを特定するために、少なくとも前記ＩＤ２と関係する前記名称ＩＤ、前記パス名称ＩＤ、前記パス階層ＩＤを対応付けた要素管理テーブルを参照し、前記ＩＤ２に対応する前記名称ＩＤ、前記パス名称ＩＤ、前記パス階層ＩＤの少なくともいずれか１つのＩＤ（以下、ＩＤ３）を求めるステップと、前記ＩＤ１と前記ＩＤ３とが一致する前記検索単位識別子のみを抽出するステップを有するプログラムを記録した可搬型媒体。

【請求項１８】　中間ノード以下を検索範囲に指定した場合における検索範囲に含まれるノードを決定する方法について、各要素実体に至るタグ名を階層順に連ねたパス名称を識別するパス名称又は、同一の親ノードを持ち同一な名称を持つタグの同一階層内での出現順序を階層順に連ねたパス階層を、１階層登り、現在位置するノードが指定した中間ノードと一致するか又は、既に検索範囲に含まれていると判定されているノードであるかを判断し、前記いずれかの条件に該当するノードである場合はそれまでたどったノード全てを検索範囲に含まれると判定し、現在位置するノードが指定した中間ノードと一致しないか又は、既に検索範囲外と判定されているノードであるかを判断し、前記いずれかの条件に該当するノードである場合はそれまでたどったノード全てを検索範囲外であると判定する処理を、最下層ノードを起点として１階層登る毎に実行し、最上位層のノードに至るまで繰り返し実行することにより検索範囲を特定する方法。

【請求項１９】　構造化文書を管理するために汎用計算機を、構造化文書の入力を行う構造化文書入力手段と、前記構造化文書入力手段により取り込んだ構造化文書を解析し該構造化文書の木構造を生成する構造解析手段と、前記構造解析手段により木構造で表現された構造化文書において、各要素実体を識別する検索単位識別子と、各要素実体に至るタグ名を階層順に連ねたパス名称を識別するパス名称ＩＤと、同一の親ノードを持ち同一な名称を持つタグの同一階層内での出現順序を階層順に連ねたパス階層を識別するパス階層ＩＤと、前記検索単位識別子から前記パス名称ＩＤと前記パス階層ＩＤを特定するために、少なくとも前記検索単位識別子と関係する前記パス名称ＩＤ及びパス階層ＩＤを対応付けた要素管理テーブルを作成する構造情報作成手段と、文字列検索を行うための文字列索引を作成する文字列索引作成手段と、検索条件の入力を行う検索条件入力手段と、前記検索条件入力手段で入力された検索条件に該当する前記パス名称ＩＤ、前記パス階層ＩＤの少なくともいずれか１つを特定する検索条件解析手段と、前記文字列索引作成手段で作成された文字列索引を用いて検索条件に該当する文字列を有する各要素実体の前記検索単位識別子を特定する文字列索引検索手段と、前記文字列索引検索手段で特定した検索単位識別子を基に前記要素管理テーブルを参照して対応するパス名称ＩＤ又はパス階層ＩＤを求め、前記パス名称ＩＤ又は前記パス階層ＩＤと前記検索条件解析手段により求めたパス名称ＩＤ又はパス階層ＩＤとが一致する検索単位識別子のみを抽出する構造照合手段、として機能させるための構造化文書管理プログラム。

【請求項２０】　各要素実体内部にさらにタグに囲まれた要素実体（子要素）を含む構造化文書の索引を作成するために汎用計算機を、各要素実体から所定の文字数で取り出した文字列が前記タグにまたがる場合は、該子要素を識別する独自の検索単位識別子を取得し、該文字列と該文字列の各文字の属する要素実体を識別する検索単位識別子と前記タグを取り除いた要素実体内での該文字列の位置を示す文字位置識別子とから成る検索用文字列索引を生成する文字列索引作成手段として機能させるための文字索引作成プログラム。

【請求項２１】　予め数値であることを定義しているタグに囲まれた文字列を含む構造化文書の索引を作成するために汎用計算機を、該タグに囲まれた文字列を識別する独自の検索単位識別子を取得し、該タグに囲まれた文字列を数値データに変換し、前記検索単位識別子と前記数値データとを対応付けた数値型索引を作成する数値型索引作成手段として機能させるための文字列索引作成プログラム。

【請求項２２】　木構造で表現されるデータにおいて所定のノード以下を検索範囲に指定した場合に、検索範囲に含まれるノードを特定するプログラムであって、各ノードが検索範囲に含まれるか否かを示す照合フラグを格納する照合テーブルを初期化する第一のステップ、参照しているノードが検索範囲内か否か又は未定であるかを、照合テーブルをもとに判断する第二のステップ、第二のステップにより検索範囲内と判断した場合は、参照しているノードについて検索範囲内を示す照合フラグを照合テーブルに設定する第三のステップ、第二のステッ

プにより検索範囲外と判断した場合は、参照しているノードについて検索範囲外を示す照合フラグを照合テーブルに設定する第四のステップ、第二のステップにより未定と判断した場合であって、さらに参照しているノードが指定したノードと一致する場合又は、既に検索範囲内である場合は、それまでたどった全てのノードについて検索範囲内を示す照合フラグを照合テーブルに設定する第五のステップ、第二のステップにより未定と判断した場合であって、さらに参照しているノードが既に検索範囲外である場合は、それまでたどった全てのノードについて範囲外を示す照合フラグを照合テーブルに設定する第六のステップ、第五のステップまたは第六のステップのいずれにも該当しない場合は、現在参照しているノードから１階層上る第七のステップ、前記第七のステップにより１階層上ったノードがルートノードである場合は、それまでたどった全てのノードについて検索範囲外を示す照合フラグを照合テーブルに設定する第八のステップ、前記第七のステップにより１階層上ったノードがルートノード以外である場合は、前記第五のステップへ戻る第八のステップ、とから構成されることにより、検索範囲を特定するプログラム。

【請求項２３】　木構造で表現される構造化文書を管理する装置であって、要素実体を識別する検索単位識別子を割当てる構造情報作成手段と、前記検索単位識別子とは別個に要素実体を特定する手段として、前記木構造において同一の親ノードを持ち同一な名称を持つタグの出現順序を階層別に連ねたパス階層を格納する手段と、前記木構造においてタグ名を階層別に連ねたパス名称を格納する手段と、を備え、さらに、前記パス階層及びパス名称と前記検索単位識別子とを関連付ける要素管理テーブルを格納する手段と、検索条件の文字列を含む要素実体の検索単位識別子を抽出する文字列索引検索手段と、文字列索引検索手段により抽出された検索単位識別子から、前記要素管理テーブルを参照し、検索条件として指定したパス階層又はパス名称を満たす文書を検索する構造照合手段と、を有する構造化文書管理装置。

【請求項２４】　木構造で表現可能なデータ構造を有するデータを管理するデータ管理装置であって、データの実体要素の特定は、前記木構造において同一の親ノードを持ち同一な名称を持つタグの出現順序を階層別に連ねたパス階層を格納する手段を用いることを特徴とするデータ管理装置。

【請求項２５】　木構造で表現されたデータのタグ名を階層別に連ねたパス名称を格納する手段をさらに備え、前記木構造におけるデータの実体要素を一意に特定するために前記パス階層を格納する手段と、前記パス名称を格納する手段とを用いることを特徴とする請求項２４記載のデータ管理装置。

【請求項２６】　同一親ノードを持ち同一のタグ名称を有する実体要素が複数存在する場合、前記パス名称は同一に表現されることを特徴とする請求項２５記載のデータ管理装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、ＳＧＭＬやＸＭＬなどの論理的な構造要素を有する構造化文書を計算機を用いて管理する文書管理システムにおける、論理構造を指定した検索を行なう構造化文書検索方法に関するものである。

【０００２】

【従来の技術】電子化文書の増大に伴い、マニュアル、議事録、仕様書等、論理的構造を有する文書を扱う構造化文書に対する関心が高まっている。それにより、文書内容のみによる検索だけでなく、構造化文書の特長を生かした、論理構造を指定した検索を行なう機能が重要となる。構造化文書はその論理構造がＤＴＤ(Ｄｏｃｕｍｅｎｔ Ｔｙｐｅ Ｄｅｆｉｎｉｔｉｏｎ：文書型定義)によって定義される。

【０００３】従来、構造化文書管理システムにおける文書の検索装置としては、特開平１０－２４０７５２号公報（以下、公知例と呼ぶ）に記載された発明が知られている。

【０００４】以下、公知例の概要について説明する。その文書登録システムの構成図は図３３に示すとおりである。公知例では登録する際、まず文書構造解析プログラム３３０１にて登録対象文書の持つ論理構造を解析して、解析済み文書データを作成し解析済み文書データ格納領域３３０５に登録する。

【０００５】次に、構造インデックス作成プログラム３３０２にて各登録対象文書の持つ論理構造を、登録順に従って順次重ね合わせ、文書中における出現位置および種別が同じである要素群は単一のメタ要素によって代表させ、文書中における出現位置が同じである文字列データ群は単一のメタ文字列データによって代表させることにより、メタ要素群およびメタ文字列データ群（公知例ではこれらを総称してメタノードと呼ぶ）の木構造から構成される構造インデックスを生成し該構造インデックスを構成する全てのメタノードに対して、それらを構造インデックスの中で一意に識別する識別子（公知例ではこれを文脈識別子と呼ぶ）を与え、構造インデックス格納領域３３０６に登録する。

【０００６】図３４は上記構造インデックスを作成する過程を示す図である。図３４において、文書１、文書２、文書３は、それぞれ登録対象文書の解析済み文書データを表わしている。これらの解析済み文書データの構造を既存の構造インデックス上に順次重ね合わせることにより、構造インデックスが形成されていく。まず最初に文書１が入力されると、最初の段階では構造インデックスは初期状態（空）であるため、該解析済みデータと等価な木構造が生成されてそのまま構造インデックスに

登録され、構造インデックスは３４０１に示す状態となる。新たに生成されたメタ要素にはＥ１からＥ５までの文脈識別子、新たに生成されたメタ文字列データにはＣ１からＣ３までの文脈識別子が割り当てられる。次に文書２が入力されると、既存の構造インデックス（３４０１）と構造が重複する部分については何も行わず、３４０１上に対応する部分がなかった部分構造（図中の網掛け部分）だけが新たに登録される。新たに生成されたメタ要素には文脈識別子Ｅ６およびＥ７、新たに生成されたメタ文字列データには文脈識別子Ｃ４が割り当てられる。次に文書３が入力されると、既存の構造インデックス（３４０２）と構造が重複する部分については何も行わず、３４０２上に対応する部分がなかった部分構造（図中の網掛け部分）だけが新たに登録される。新たに生成されたメタ要素には文脈識別子Ｅ８、Ｅ９およびＥ１０、新たに生成されたメタ文字列データには文脈識別子Ｃ５およびＣ６が割り当てられる。このようにして、３個の文書が登録された段階で、構造インデックスは３４０３に示す状態となる。

【０００７】次に、構造化全文データ生成プログラム３３０３にて各登録対象文書について、その文書に対応する解析済み文書データ中に含まれるすべての文字列と、その文字列を構造インデックス中で示される文脈識別子との対応関係の定義から構成されるデータ（公知例ではこれを構造化全文データと呼ぶ）を生成し、構造化全文データ格納領域３３０７に登録する。

【０００８】次に、文字列インデックス作成プログラム３３０４にて、各登録対象文書に対応する構造化全文データから、前記文脈識別子を含んだ全文検索を行なうための文字列インデックスを作成し、文字列インデックス格納領域３３０８に登録する。

【０００９】図３５は、文字列インデックスの例を示したものであり、部分文字列（３４０４）を２文字とした場合の例を示している。各部分文字列に対して該部分文字列を含む文書を識別する文書識別子（３４０５）、該文書中において前記部分文字列を含む文字列データの文書構造中における位置を識別する文脈識別子（３４０６）、文書中における前記部分文字列の文字位置（３４０７）から構成されている。なお、図中の“Ｘ”は文字列の直前に位置する文字の位置を“Ｘ”として相対的な文字位置を示している。

【００１０】また、公知例における検索は、まず前記構造インデックスを参照し、指定された構造条件を満たす文脈識別子の集合を決定する。

【００１１】次に、それらの文脈識別子をキーとして文字列を検索することにより、指定条件を満たす文書群を求める。

【００１２】また、公知例における登録の際に、例えば強調表示のような非構造的要素（Ｍｉｘｅｄ　Ｃｏｎｔｅｎｔと呼ぶ：詳細は実施の形態３で説明する）が含まれる場合、該構造を無視して文字列インデックスを作成する。

【００１３】

【発明が解決しようとする課題】上記従来技術の方法では、図３５に示すように全文検索を行なうための文字列インデックス内に、登録文書を識別する文書識別子と、論理構造に関する情報である文脈識別子と、文字連鎖の位置を示す文字位置という３要素のデータを含んでいるため、前記文字列インデックスのサイズが大きくなり、そのためメモリ量が増大し、装置のコストアップにつながるという課題を有していた。

【００１４】また上記従来技術の方法では、図３５に示すように文字列インデックス内の各文字連鎖に論理構造に関する情報である文脈識別子を含んでいるため、複数の登録文書の１つについて要素実体を追加、変更したことにより、複数の登録文書の論理構造を順次重ね合わせることによって形成される構造インデックス（図３４）が変化した場合、文字列インデックスの文脈識別子を更新する必要が発生し、要素実体の文字連鎖数が膨大の場合、処理量も膨大になるという課題を有していた。

【００１５】以下、この課題について具体例を通して詳細に説明する。

【００１６】図３６は２つの文書が登録されている場合の例で、このうち１つの登録文書を変更する例を示している。この例では、文書１と文書２の論理構造は同一であるので、作成される構造インデックスの論理構造も文書１又は２と同一である。この例では、文書２の第１章と第２章の間に新たな章を追加して３つの章から成る文書に変更する例を示している。すなわち文書２に新たに第２章となるブロック（図３６の４０００）を追加する例を示している。このとき、変更前に第２章であったブロック（図３６の４００１）が第３章となるが、変更前の構造インデックスには、文書１、２とも第２章までしかなかった為、文書２の第３章に相当する文脈識別子は存在しない（図３６の変更前構造インデックス）。そこで、図３６（変更後構造インデックス）に示すように構造インデックスを更新する必要がある。

【００１７】図３６の更新後の構造インデックスに示すように、文書２で新たに第３章となった要素実体に対応する文脈識別子は‘Ｃ４’となっている。しかし変更前、前記要素実体に対応する文脈識別子は‘Ｃ３’であったので、前記要素実体の文字列インデックスに保持されている各文字連鎖の文脈識別子を‘Ｃ３’から‘Ｃ４に変更する必要がある。例えば、文書２の第３章に相当する要素実体が１００文字から構成されているとすると、２文字連鎖で索引を作成する場合、９９個の文字連鎖について文脈識別子を変更する必要がある。このように要素実体の文字連鎖数に応じて処理量も大きくなってしまうという課題を有していた。

【００１８】なお、変更後に第２章となった要素実体に

新たな文脈識別子‘Ｃ４’を付与し、変更前第２章で変更後第３章となる要素実体にはそのままの文脈識別子‘Ｃ３とする更新方法も考えられるが、この場合は文書１の第２章に相当する要素実体の文字列インデックスの文字連鎖について、文脈識別子を‘Ｃ３’から‘Ｃ４’に変更する必要がある。この例では登録文書が２つなので、上述の方法と更新にかかる処理量は同一であるが、登録文書の数が増加した場合、第２章を有する全ての登録文書の要素実体について、その文字列インデックスを文脈識別子を‘Ｃ３’から‘Ｃ４’に変更する必要があるため、かえって処理量が増加してしまう結果になる。

【００１９】また別の課題として、従来技術の構造インデックスは図３４に示すように複数の登録文書の論理構造を順次重ね合わせることによって形成されるので、登録文書の論理構造がほぼ同一の場合は新たに文脈識別子を付与する機会は少ないが、各登録文書の論理構造が大きく異なる場合は論理構造の重なりが少なくなり、このような論理構造が異なる登録文書が膨大に登録された場合は、文脈識別子の数が膨大になるという課題を有していた。

【００２０】また従来技術の構造インデックスは、図３４に示すように複数の登録文書の論理構造を順次重ね合わせることによって形成されるので、この方法により形成される構造インデックスには、１つの親ノードから同一のタグ名を有する子ノードが複数出ている構造も発生する場合がある。このとき検索範囲として或るタグ名を指定した場合、各ノードのタグ名が該当するタグ名であるか否かをチェックする必要があるが、たとえ上記のように１つの親ノードから同一のタグ名を有する子ノードが複数出ていたとしても、各子ノードの１つ１つについて該当するタグ名を有するノードであるか否かをチェックするＯＲ検索が必要の為、検索が遅くなるという課題を有していた。

【００２１】また上記従来の方法では、要素実体である“段落”要素中にＭｉｘｅｄ　Ｃｏｎｔｅｎｔとして“キーワード”要素を含むような場合、“キーワード”タグの構造を無視して文字列インデックスを作成するため、「“キーワード”タグの中に“○○”を含む文書」というような検索条件に対応できないという課題を有していた。

【００２２】本発明は上記従来技術の課題を解決するもので、構造化文書を対象とした全文検索において、様々な論理構造指定検索に対応すること、さらに検索用索引のサイズ削減、文書の一部変更・一部削除時における検索用索引の変更作業の簡易化、中間ノード以下を指定した高速な検索、そしてＭｉｘｅｄ　Ｃｏｎｔｅｎｔにまたがる検索、およびＭｉｘｅｄ　Ｃｏｎｔｅｎｔである要素を指定した検索を行なうことを目的とする。

【００２３】

【課題を解決するための手段】上記課題を解決するために、請求項１では各要素実体を識別する検索単位識別子と、各要素実体の前記木構造における位置を表現した要素実体位置識別子と、前記検索単位識別子から前記要素実体位置識別子を特定するために、少なくとも前記検索単位識別子と関係する前記要素実体位置識別子を対応付けた要素管理テーブルを作成する構造情報作成手段を有することにより、登録文書の構造が変化した場合でも前記要素管理テーブルを更新するのみで対応が可能となり、従来技術のように文書構造が変化する度に文字列インデックス内の文脈識別子を変更する必要はないので、登録文書の論理構造の変化する度に文字列インデックス更新のための膨大な処理量が発生することはない。

【００２４】請求項２では各要素実体を識別する検索単位識別子と、各要素実体に至るタグ名を階層順に連ねたパス名称を識別するパス名称ＩＤと、同一の親ノードを持ち同一な名称を持つタグの同一階層内での出現順序を階層順に連ねたパス階層を識別するパス階層ＩＤと、前記検索単位識別子から前記パス名称ＩＤと前記パス階層ＩＤを特定するために、少なくとも前記検索単位識別子と関係する前記パス名称ＩＤ及びパス階層ＩＤを対応付けた要素管理テーブルを作成する構造情報作成手段を有することにより、登録文書の構造が変化した場合でも前記要素管理テーブルを更新するのみで対応が可能となり、従来技術のように登録文書の論理構造の変化する度に文字列インデックス更新のための膨大な処理量が発生することはない。また、パス名称ＩＤとパス階層ＩＤを導入することにより、従来技術のように検索範囲を特定する際のＯＲ検索が不要になる。また、登録文書の論理構造が異なる文書を多く登録する場合でも、要素実体をパス名称ＩＤとパス階層ＩＤとで特定するので、従来技術のように複数の登録文書の論理構造を順次重ね合わせることによって形成される場合に必要となる文脈識別子数よりは少なくて済む。

【００２５】請求項３ではタグ名を識別する名称ＩＤと、各要素実体を識別する検索単位識別子と、前記検索単位識別子から前記名称ＩＤを特定するために、少なくとも前記検索単位識別子と関係する前記名称ＩＤを対応付けた要素管理テーブルを作成する構造情報作成手段を有することにより、検索範囲として登録文書のノードのタグ名を指定することが可能となる。

【００２６】請求項４では文字列検索結果一覧や各要素実体表示のためのデータを作成する結果作成手段と、前記結果作成手段で作成された検索結果を端末に表示する結果表示手段とを有することにより、使用者に検索結果を表示することが可能となる。

【００２７】請求項５ではネットワーク上に、構造化文書の入力を行う構造化文書入力手段と、前記構造化文書入力手段により取り込んだ構造化文書を解析し該構造化文書の木構造を生成する構造解析手段と、前記構造解析手段により木構造で表現された構造化文書において、各

要素実体を識別する検索単位識別子と、各要素実体に至るタグ名を階層順に連ねたパス名称を識別するパス名称ＩＤと、同一の親ノードを持ち同一な名称を持つタグの同一階層内での出現順序を階層順に連ねたパス階層を識別するパス階層ＩＤと、前記検索単位識別子から前記パス名称ＩＤ及び前記パス階層ＩＤを特定するために、少なくとも前記検索単位識別子と関係する前記パス名称ＩＤ及びパス階層ＩＤを対応付けた要素管理テーブルを作成する構造情報作成手段とから成る構造化文書登録部を独立して設けることにより、ネットワークを介して遠隔から構造化文書の登録をすることが可能となり、登録文書の構造が変化した場合でも前記要素管理テーブルを更新するのみで対応が可能となり、従来技術のように登録文書の論理構造の変化する度に文字列インデックス更新のための膨大な処理量が発生することはない。またパス名称ＩＤとパス階層ＩＤを導入することにより、従来技術のように検索範囲を特定する際のＯＲ検索が不要になる。また登録文書の論理構造が異なる文書を多く登録する場合でも、要素実体をパス名称ＩＤとパス階層ＩＤとで特定するので、従来技術のように複数の登録文書の論理構造を順次重ね合わせることによって形成される場合に必要となる文脈識別子数よりは少なくて済む。

【００２８】請求項６及び７では構造化文書の木構造が変化した場合に、要素管理テーブルに記録されたパス名称ＩＤ、パス階層ＩＤのうち、変更が必要なＩＤを更新することにより、登録文書の構造が変化した場合でも前記要素管理テーブルを更新することで対応が可能となり、従来技術のように登録文書の論理構造の変化する度に文字列インデックス更新のための膨大な処理量が発生することはない。

【００２９】請求項８ではネットワーク上に、構造化文書の入力を行う構造化文書入力手段と、前記構造化文書入力手段により取り込んだ構造化文書を解析し該構造化文書の木構造を生成する構造解析手段と、前記構造解析手段により生成された木構造からタグ名を識別する名称ＩＤと、各要素実体を識別する検索単位識別子と、前記検索単位識別子から前記名称ＩＤを特定するために、少なくとも前記検索単位識別子と関係する前記名称ＩＤを対応付けた要素管理テーブルを作成する構造情報作成手段とから成る構造化文書登録部を独立して設けることにより、ネットワークを介して遠隔から構造化文書の登録が可能となる。

【００３０】請求項９では各要素実体から所定の文字数で取り出した文字列が前記タグにまたがる場合は、該子要素を識別する独自の検索単位識別子を取得し、該文字列と該文字列の各文字の属する要素実体を識別する検索単位識別子と前記タグを取り除いた要素実体内での該文字列の位置を示す文字位置識別子とから成る検索用文字列索引を生成する文字列索引作成部により、Ｍｉｘｅｄ Ｃｏｎｔｅｎｔを含んだ構造化文書でも検索が可能となる。また作成される文字索引は前記検索単位識別子と前記文字位置識別子の２要素から成るので、従来技術では３要素から成る文字列インデックスと比べメモリ量を削減することができ、装置のコストダウンを実現することができる。

【００３１】請求項１０では予め数値であることを定義しているタグに囲まれた文字列を識別する独自の検索単位識別子を取得し、該タグに囲まれた文字列を数値データに変換し、前記検索単位識別子と前記数値データとを対応付けた数値型索引を生成する数値型索引作成手段により、特定の数値範囲を指定した検索が可能になる。

【００３２】請求項１１ではネットワーク上に、タグ名を識別する名称ＩＤと、各要素実体に至るタグ名を階層順に連ねたパス名称を識別するパス名称ＩＤと、同一の親ノードを持ち同一な名称を持つタグの同一階層内での出現順序を階層順に連ねたパス階層を識別するパス階層ＩＤと、各要素実体を識別する検索単位識別子と、前記検索単位識別子から前記名称ＩＤを特定するために、少なくとも前記検索単位識別子と関係する前記名称ＩＤを対応付けた要素管理テーブルまたは、前記検索単位識別子から前記パス名称ＩＤと前記パス階層ＩＤを特定するために、少なくとも前記検索単位識別子と関係する前記パス名称ＩＤ及びパス階層ＩＤを対応付けた要素管理テーブルの少なくともいずれか一方を記憶するデータ格納部と、検索条件の入力を行う検索条件入力手段と、前記検索条件入力手段で入力された検索条件から検索条件に該当する前記名称ＩＤ、前記パス名称ＩＤ、前記パス階層ＩＤの少なくともいずれか１つ（ＩＤ１）を特定する検索条件解析手段と、検索条件に該当する文字列を有する前記検索単位識別子を求める文字列索引検索手段と、前記文字列索引検索手段で特定した検索単位識別子を基に前記要素管理テーブルを参照して対応する名称ＩＤ、パス名称ＩＤ、パス階層ＩＤの少なくともいずれか１つ（ＩＤ２）を求め、前記ＩＤ２と前記検索条件解析手段により求めたＩＤ１とが一致する検索単位識別子のみを抽出する構造照合手段を備えた文字列検索部をそれぞれ独立して設けることにより、ネットワークを介して遠隔からの文字列検索が可能となる。

【００３３】請求項１２では予め数値であることを定義しているタグに囲まれた文字列を含む構造化文書の数値範囲検索において、前記タグに囲まれた文字列を識別する独自の検索単位識別子と前記タグに囲まれた文字列を数値に変換した数値データとを対応付けた数値型索引を参照し、検索条件に該当する前記検索単位識別子を抽出する数値型索引検索手段を有することを特徴とする請求項１１記載の文字列検索部を有していることにより、ネットワークを介して遠隔から、指定した範囲の数値を有する要素実体の検索単位識別子を求めることが可能となる。

【００３４】請求項１３では構造化文書を読み込むステ

ップと、各要素実体に至るタグ名を階層順に連ねたパス名称を識別するパス名称ＩＤと、同一の親ノードを持ち同一な名称を持つタグの同一階層内での出現順序を階層順に連ねたパス階層を識別するパス階層ＩＤを生成するステップと、要素実体を有するか否かを判断するステップと、各要素実体を識別する検索単位識別子を生成するステップと、前記検索識別子から前記パス名称ＩＤ及び前記パス階層ＩＤを特定するために、少なくとも前記検索単位識別子と関係する前記パス名称ＩＤ及びパス階層ＩＤを対応付けた要素管理テーブルを作成するステップを有するプログラムを記録した可搬型媒体により、汎用計算機に上記プログラムをインストールすることで、構造化文書を登録する構造化文書登録部の機能を持たせることが可能となる。

【００３５】請求項１４では構造化文書を読み込むステップと、タグ名を識別する名称ＩＤを生成するステップと、要素実体を有するか否かを判断するステップと、各要素実体を識別する検索単位識別子を生成するステップと、前記検索単位識別子から前記名称ＩＤを特定するために、少なくとも前記検索単位識別子と関係する前記名称ＩＤを対応付けた要素管理テーブルを作成するステップを有するプログラムを記録した可搬型媒体により、汎用計算機に上記プログラムをインストールすることで、構造化文書を登録する構造化文書登録部の機能を持たせることが可能となる。

【００３６】請求項１５では、要素実体内部にさらにタグに囲まれた要素実体（子要素）を有する構造化文書の文字索引の生成方法について、構造解析済みデータを読み込むステップと、要素実体を有するか否かをチェックするステップと、要素実体を識別するための検索単位識別子を取得するステップと、前記子要素を含むか否かを調べるステップと、該子要素を識別する検索単位識別子を取得するステップと、要素実体から１以上の所定文字数を単位とする文字列を取り出すステップと、前記文字列の各文字の属する検索単位識別子を求めるステップと、該文字列及び該文字列の各文字の属する前記検索単位識別子及びタグを取り除いた要素実体内での当該文字列の位置を示す文字位置識別子を有する検索文字列索引を生成するステップとを有するプログラムを記録した可搬型媒体により、汎用計算機に上記プログラムをインストールすることにより、Ｍｉｘｅｄ　Ｃｏｎｔｅｎｔを含んだ構造化文書でも検索が可能な文字列索引を作成する文字列索引作成部の機能を持たせることが可能となる。

【００３７】請求項１６では、構造化文書の数値検索用索引生成方法について、構造化文書を読み込むステップと、予め数値であることを定義しているタグに囲まれた文字列であるか否かを判断するステップと、数値であることを定義したタグに囲まれた文字列を識別するための検索単位識別子を取得するステップと、該文字列を数値に変換するステップと、前記検索単位識別子と前記数値とからなる数値型索引を生成するステップを有するプログラムを記録した可搬型媒体により、汎用計算機に上記プログラムをインストールすることにより、数値範囲を指定した検索も可能な文字列索引を生成する文字列索引作成部の機能を持たせることが可能となる。

【００３８】請求項１７では、構造化文書の検索方法について、検索条件を読み込むステップと、前記検索条件に該当するタグ名を識別する名称ＩＤ又は、各要素実体に至るタグ名を階層順に連ねたパス名称を識別するパス名称ＩＤ又は、同一の親ノードを持ち同一な名称を持つタグの同一階層内での出現順序を階層順に連ねたパス階層を識別するパス階層ＩＤのいずれかのＩＤ（以下、ＩＤ１）に変換するステップと、検索条件に該当する文字列を有する各要素実体を識別する検索単位識別子（以下、ＩＤ２）を特定するステップと、前記ＩＤ２から前記名称ＩＤ、前記パス名称ＩＤ、前記パス階層ＩＤを特定するために、少なくとも前記ＩＤ２と関係する前記名称ＩＤ、前記パス名称ＩＤ、前記パス階層ＩＤを対応付けた要素管理テーブルを参照し、前記ＩＤ２に対応する前記名称ＩＤ、前記パス名称ＩＤ、前記パス階層ＩＤの少なくともいずれか１つのＩＤ（以下、ＩＤ３）を求めるステップと、前記ＩＤ１と前記ＩＤ３とが一致する前記検索単位識別子のみを抽出するステップを有するプログラムを記録した可搬型媒体により、汎用計算機に上記プログラムをインストールすることにより、文字列検索部の機能を持たせることが可能となる。

【００３９】請求項１８では、中間ノード以下を検索範囲に指定した場合における検索範囲に含まれるノードを決定する方法について、各要素実体に至るタグ名を階層順に連ねたパス名称を識別するパス名称又は、同一の親ノードを持ち同一な名称を持つタグの同一階層内での出現順序を階層順に連ねたパス階層を、１階層登り、現在位置するノードが指定した中間ノードと一致するか又は、既に検索範囲に含まれていると判定されているノードであるかを判断し、前記いずれかの条件に該当するノードである場合はそれまでたどったノード全てを検索範囲に含まれると判定し、現在位置するノードが指定した中間ノードと一致しないか又は、既に検索範囲外と判定されているノードであるかを判断し、前記いずれかの条件に該当するノードである場合はそれまでたどったノード全てを検索範囲外であると判定する処理を、最下層ノードを起点として１階層登る毎に実行し、最上位層のノードに至るまで繰り返し実行することにより検索範囲を特定する方法により、検索範囲として或る中間ノード以下を指定した場合に検索範囲に含まれるノードを特定することが可能となる。

【００４０】請求項１９の発明は、請求項２に記載の構造化文書管理装置を、汎用計算機とプログラムによって実現することを可能とするものである。

【００４１】請求項２０の発明は、請求項９に記載の文字列索引作成装置を、汎用計算機とプログラムによって実現することを可能とするものである。
【００４２】請求項２１の発明は、請求項１０に記載の文字列索引作成装置を、汎用計算機とプログラムによって実現することを可能とするものである。
【００４３】請求項２２の発明は、木構造を有するデータを検索するために、検索範囲として所定のノード以下を指定した場合に、各ノードが検索範囲に含まれるか否かを示す照合テーブルを順次作成していくプログラムにより、検索範囲の特定を効率良く実現するものである。
【００４４】請求項２３の発明は、木構造で表現される構造化文書を管理する装置であって、要素実体を識別する検索単位識別子を割当てる構造情報作成手段と、前記検索単位識別子とは別個に要素実体を特定する手段として、前記木構造において同一の親ノードを持ち同一な名称を持つタグの出現順序を階層別に連ねたパス階層を格納する手段と、前記木構造においてタグ名を階層別に連ねたパス名称を格納する手段と、を備え、さらに、前記パス階層及びパス名称と前記検索単位識別子とを関連付ける要素管理テーブルを格納する手段と、検索条件の文字列を含む要素実体の検索単位識別子を抽出する文字列索引検索手段と、文字列索引検索手段により抽出された検索単位識別子から、前記要素管理テーブルを参照し、検索条件として指定したパス階層又はパス名称を満たす文書を検索する構造照合手段と、を有する構造化文書管理装置であって、効率良く文書検索を実現することが可能となる。
【００４５】請求項２４の発明は、木構造で表現可能なデータ構造を有するデータを管理するデータ管理装置であって、データの実体要素の特定は、前記木構造において同一の親ノードを持ち同一な名称を持つタグの出現順序を階層別に連ねたパス階層を格納する手段を用いることを特徴とするデータ管理装置であって、木構造で表現できるデータの管理を少ない個数のＩＤにより管理することが可能になる。
【００４６】請求項２５の発明は、木構造で表現されたデータのタグ名を階層別に連ねたパス名称を格納する手段をさらに備え、前記木構造におけるデータの実体要素を一意に特定するために前記パス階層を格納する手段と、前記パス名称を格納する手段とを用いることを特徴とする請求項２４記載のデータ管理装置であって、木構造で表現できるデータをパス階層及びパス名称で特定することにより少ない個数のＩＤにより管理することが可能になる。
【００４７】請求項２６の発明は、同一親ノードを持ち同一のタグ名称を有する実体要素が複数存在する場合、前記パス名称は同一に表現されることを特徴とする請求項２５記載のデータ管理装置であって、データの検索においていわゆるＯＲ検索が不要となり、高速に検索することが可能にすることが可能となる。
【００４８】
【発明の実施の形態】以下、本発明の実施の形態について説明する。なお、本発明はこれら実施の形態に何ら限定されるものではなく、その要旨を逸脱しない範囲において、種々なる態様で実施し得る。
【００４９】（実施の形態１）図１は本発明の実施の形態１における構造化文書管理装置の構成図である。図１に示す構造化文書管理装置は、端末１０１、構造化文書入力手段１０２、検索条件入力手段１０３、結果表示手段１０４、検索エンジン１０５、データ格納部１０６からなる。
【００５０】端末１０１は、文書検索における検索条件の指定および検索結果の表示に使用する。
【００５１】構造化文書入力手段１０２は、登録対象文書を格納しておき、文書の登録を行なう際にここから検索エンジン１０５へデータを送る。検索条件入力手段１０３は、端末１０１から入力された検索条件を検索エンジン１０５へ送る。
【００５２】結果表示手段１０４は、検索結果を検索エンジン１０５から受け取り、端末１０１に表示する。
【００５３】検索エンジン１０５は、実際に構造化文書の登録、検索および検索結果の作成を行なう。まず、登録に関して、１０７は登録対象文書の論理構造を解析する構造解析手段、１０８は前記構造解析手段によって論理構造に分けられた各要素の論理構造に関する情報を作成する構造情報作成手段、１０９は文字列に対して高速に検索を行なうための文字列索引を作成する文字列索引作成手段である。これら１０７、１０８、および１０９についての詳細は、文書登録処理の流れの説明の中で述べる。次に検索に関して、１１０は検索条件入力手段１０３から受けた検索条件中の論理構造に関する条件を、本検索エンジン内における構造条件の表現方法に変換する検索条件解析手段、１１１は前記文字列索引を用いて検索条件中の検索文字列で検索処理を行なう文字列索引検索手段、１１２は前記文字列索引検索手段で得られた文字列検索結果群の中から、前記検索条件解析手段で変換した本検索エンジン内における構造条件に一致するものだけを抽出する構造照合手段である。１１０、１１１、および１１２についての詳細は文書検索の流れの説明の中で述べる。次に結果作成に関して、１１３は検索結果の一覧や、実体表示のためのデータを作成し結果表示手段１０４へ渡す結果作成手段である。
【００５４】データ格納部１０６は、構造解析手段１０７によって作成された構造解析済みデータを格納する構造解析済みデータ格納手段１１４、文書中の検索対象要素ごとに論理構造情報を格納した要素管理テーブル格納手段１１５、最上位階層から順にタグ名を連ねて記述した文字列（以下、パス名称と呼ぶ）を管理し、各パス名称にＩＤを割当てたパス名称インデックスを格納するパ

ス名称インデックス格納手段１１６、パス名称の各階層の出現順序（同じ親要素を持つ同じタグ名の要素の中で何番目に出現した要素かを示す番号）を連ねて記述した文字列（以下、パス階層と呼ぶ）を管理し、各パス階層にＩＤを割当てたパス階層インデックスを格納するパス階層インデックス格納手段１１７、各要素のタグ名に対してＩＤを割当てた名称ＩＤテーブルを格納する名称ＩＤテーブル格納手段１１８、前記文字列索引作成手段１０９によって作成された文字列索引を格納する文字列索引格納手段１１９、登録文書の実体データを格納する実体データ格納手段１２０、検索結果一覧のためのデータを格納する一覧データ格納手段１２１からなり、構造化文書の検索および結果表示に用いるデータの格納に使用する。

【００５５】次に本実施の形態における文書登録の処理を具体的な構造化文書の例を用いて説明する。

【００５６】まず、構造化文書入力手段１０２から登録対象文書を読み込む。次に構造解析手段１０７によって登録対象文書の構造を理解できる形に変換する。この構造解析手段１０７によって、文字の並びとしての構造化文書が構造情報作成手段１０８に理解できるデータ構造に変換され（以下、構造解析済みデータと呼ぶ）、構造解析済みデータ格納手段１１４に格納される。

【００５７】次に構造情報作成手段１０８で、前記構造解析手段によって論理構造に分けられた各要素の論理構造に関する情報を作成する。

【００５８】図２は構造化文書の一例である。図２の構造化文書を構造解析手段１０７によって解析した結果得られる木構造は図３のようになる。以下ではこの論理構造を持った構造化文書を中心に説明する。図３において実体（テキスト）を持つ要素（以下、要素実体）は網掛けで表示されている。またこれら要素実体は、検索エンジン内で検索単位を一意に表す符号（以下、検索単位識別子と呼ぶ）が割当てられる。この検索単位識別子は対象とする文書内の論理的位置とは無関係な符号であり、例えば、数値であっても良い。

【００５９】図３において要素実体の下段に書かれた数値が検索単位識別子の例である。また、要素実体は上述のパス名称インデックス、パス階層インデックス、名称ＩＤのいずれか１つ又は上記インデックスの組み合わせにより特定が可能なので、上記３種のインデックスを総称して「要素実体位置識別子」という。

【００６０】図４は構造情報作成手段１０８の処理の流れである。まず、登録対象文書の構造解析済みデータを構造解析済みデータ格納手段１１４から読込み、登録対象文書ごとに一意な番号（以下、文書番号と呼ぶ）を割当てる（ステップ４０１）。

【００６１】次に登録対象文書の各要素に対して以下の処理を繰り返す。まず、現在参照中の要素の名称ＩＤの取得を行なう（ステップ４０２）。図５は図３のような木構造を持つ構造化文書を登録した時に最終的に作成される名称ＩＤテーブルの例である。図３の３０１の要素のタグ名は“段落”であるから、図５より名称ＩＤは“Ｔ９”である。ステップ４０２では、この名称ＩＤテーブルに現在参照中の要素に該当するタグ名と名称ＩＤのレコードが存在する場合はその名称ＩＤを取得し、存在しない場合にはそのタグ名と名称ＩＤのレコードを新たに作成し、名称ＩＤテーブル格納手段１１８に格納するとともに、その名称ＩＤを取得する。次に現在参照中の要素のパス名称ＩＤの取得を行なう（ステップ４０３）。図６は図３のような木構造を持つ構造化文書を登録した時に、最終的に作成されるパス名称インデックスの例である。パス名称インデックスは、登録対象文書のパス名称に一意なＩＤ（パス名称ＩＤ）を割当てたものである。また各パス名称ＩＤは最下層のタグ名の名称ＩＤの情報を持つ。図３の３０１の要素のパス名称は「／論文／本文／章／節／段落」であり、このパス名称に割当てられたパス名称ＩＤは、図６の例では６０１に示される値（Ｎ１１）である。ステップ４０３では、このパス名称インデックスに現在参照中の要素に該当するパス名称のノードが存在する場合はそのパス名称ＩＤを取得し、存在しない場合にはそのパス名称のノードとそのパス名称ＩＤを新たに作成し、パス名称インデックス格納手段１１６に格納するとともに、そのパス名称ＩＤを取得する。なお、ここでパス名称を表現する際に、各階層の区切り文字として“／”（スラッシュ）を用いたが、これはタグ名に用いられない文字である限りどのようなものでも構わない。次に現在参照中の要素のパス階層ＩＤの取得を行なう（ステップ４０４）。図７は図３のような木構造を持つ構造化文書を登録した時に、最終的に作成されるパス階層インデックスの例である。パス階層インデックスは、登録対象文書のパス階層に一意なＩＤ（パス階層ＩＤ）を割当てたものである。図３の３０１の要素のパス階層は「／１／１／１／１／２」であり、このパス階層に割当てられたパス階層ＩＤは図７の例では７０１に示される値（Ｌ５）である。ステップ４０４では、このパス階層インデックスに現在参照中の要素に該当するパス階層のノードが存在する場合はそのパス階層ＩＤを取得し、存在しない場合にはそのパス階層のノードとそのパス階層ＩＤを新たに作成し、パス階層インデックス格納手段１１７に格納するとともに、そのパス階層ＩＤを取得する。なお、ここでパス階層を表現する際に、パス名称と同様に各階層の区切り文字として“／”（スラッシュ）を用いたが、これは出現順序を表す数字に用いられない文字である限りどのようなものでも構わない。次に現在参照中の要素が実体を持つかどうかチェックし（ステップ４０５）、実体を持たない場合はステップ４０８へ進む。実体を持つ場合、ステップ４０６へ進み、この要素に検索単位識別子を割当てる。次にステップ４０７で要素管理テーブルに現在参照中の要

素のレコードを追加する。図８は要素管理テーブルの例であり、８０１は図３の３０１の要素に関するレコードに該当する。実施の形態１における要素管理テーブルは、検索単位識別子をキーとして文書番号、パス名称ＩＤ、パス階層ＩＤ、名称ＩＤを管理する。次にステップ４０８で登録対象文書の全要素についてステップ４０２から４０７の処理を終了したか調べ、まだ未処理の要素が存在したらステップ４０２以降の処理を繰り返す。

【００６２】次に文字列索引作成手段１０９では、各要素ごとに要素内容の検索用文字列索引を作成する。文字列索引作成手段１０９の処理の流れを図９を用いて説明する。

【００６３】まず構造解析済みデータ格納手段１１４から登録対象文書の構造解析済みデータを読み込む（ステップ９０１）。次に現在参照中の要素が実体を持つかどうかチェックし（ステップ９０２）、実体を持たない場合はステップ８０７へ進む。実体を持つ場合、ステップ９０３へ進み、構造情報作成手段１０８の処理ステップ４０６でこの要素に割当てた検索単位識別子を取得する。次に該要素内容の文字列についてあらかじめ定めた文字数の文字連鎖を取り出す（ステップ９０４）。

【００６４】この文字連鎖について、該当する検索単位識別子、および該文字連鎖先頭文字がその要素内容において何番目の文字かを表す番号（以下、文字位置番号と呼ぶ）の情報を文字列索引に追加する（ステップ９０５）。ステップ９０４、９０５の処理を該要素の全文字列について繰り返す（ステップ９０６）。次にステップ９０７で登録対象文書の全要素についてステップ９０２から９０６の処理を終了したか調べ、まだ未処理の要素が存在したらステップ９０２以降の処理を繰り返す。

【００６５】全要素についてステップ９０２から９０６の処理を終了したら、最後にここで作成した文字列索引を文字列索引格納手段１１９に追加する（ステップ９０８）。

【００６６】図１０は文字列索引作成手段１０９によって図２の構造化文書のうち３行目の「<;タイトル>;　構造化文書管理　<;／タイトル>;」という要素について作成した文字列索引の例の一部を示した図である。図１０の１００１は「検索単位識別子が“１”の要素の文字列中に“構造”という文字連鎖が先頭から“１”文字目の位置から存在する」ということを表している。なお、図１０は文字列索引の一部しか示していないが、実際は登録対象文書の全要素の全文字列について文字列索引が作成される。

【００６７】なお、この例では２文字ずつ文字連鎖を取り出してそれぞれに文字列索引を作成しているが、この文字連鎖は２文字ずつでなくても構わない。また、以上の登録処理を登録対象文書が入力されるごとに繰り返すことにより、構造情報と文字列索引が追加されてゆく。

【００６８】なお、図５他において名称ＩＤ、パス名称ＩＤおよびパス階層ＩＤは“Ｔ９”や“Ｎ１１”や“Ｌ５”といった文字で表現しているが、これらはそれぞれ、名称（タグ名）を一意に特定するＩＤ、パス名称を一意に特定するＩＤ、パス階層を一意に特定するＩＤであればどのようなものでも構わない。次に本実施の形態における文書検索の処理の流れを具体例を示して説明する。

【００６９】なお、以下に示す本実施の形態における文書検索処理の説明においては、名称ＩＤテーブル、パス名称インデックス、パス階層インデックス、要素管理テーブルには、それぞれ図５、図６、図７、図８のようなデータが格納されているものとして説明を行なう。

【００７０】まず検索条件入力手段１０３を通して、端末１０１から「パス名称が“／論文／書誌／タイトル”である要素に、“構造化”という文字列が含まれる文書」という条件が与えられたとする。

【００７１】図１１は検索条件解析手段１１０の処理の流れを示した図である。ここでの例は、検索条件の構造指定としてパス名称のみ指定されているので、図１１のＣａｓｅ３に該当する。Ｃａｓｅ３ではステップ１１０２で、パス名称インデックス格納手段１１６に格納されているパス名称インデックスを参照し、検索条件のパス名称をパス名称ＩＤに変換する。パス名称インデックスが図６の場合、検索条件のパス名称“／論文／書誌／タイトル”は、パス名称ＩＤ“Ｎ２”に変換される。

【００７２】次に文字列索引検索手段１１１で、検索条件の文字列について検索処理を行なう。図１２は文字列索引検索手段１１１での処理を図に示したものである。ここでの例では検索条件の文字列は“構造化”であり、これは２文字ずつの文字連鎖として“構造”と“造化”が取り出せる。ここで取り出す文字連鎖の文字数は、文字列索引作成手段１０９で作成する文字連鎖の文字数と同一とする。この２つの文字連鎖について図１２の１２１０に示すような文字列索引が作成されているとして、この中から検索単位識別子が同一で、かつ“構造”の文字連鎖から“造化”の文字連鎖に対して文字位置番号が連続しているものを文字列索引検索手段１１１の結果として抽出する。図１２の例では検索単位識別子が同一なものとして１２２１、１２２２、１２２３を取り出すことが出来る。更にその中で文字位置番号が連続しているのは１２２１と１２２３であり、これらの検索単位識別子を抽出する。

【００７３】次に構造照合手段１１２で、文字列索引検索手段１１１で得られた検索単位識別子群の中から、検索条件の構造指定を満たす最終的な検索結果を求める。図１３は、構造照合手段１１２の処理の流れを示した図である。図１３におけるＣａｓｅ１からＣａｓｅ４は、図１１の検索条件の構造指定パターンＣａｓｅ１からＣａｓｅ４と同様である。ここでの例ではＣａｓｅ３（パス名称のみ指定）であるので、ステップ１３０３でパス

名称の照合を行なう。図１４はこの例における構造照合処理の詳細を示す図である。まず文字列索引検索手段１１１で得られた検索単位識別子（１４０１）をキーとして要素管理テーブルを参照する。そこで該検索単位識別子のパス名称ＩＤが、検索条件解析手段１１０で求めた検索条件のパス名称ＩＤ（この例では“Ｎ２”）と一致するものだけを最終的な検索結果とする。

【００７４】なお、本実施の形態では検索条件の構造指定として、タグ名を指定した検索（Ｃａｓｅ１）、タグ名とその出現順序を指定した検索（Ｃａｓｅ２）、パス名称とパス階層を指定した検索（Ｃａｓｅ４）にも対応可能である。以下でそれぞれＣａｓｅでの処理について簡潔に説明する。

【００７５】タグ名を指定した検索（Ｃａｓｅ１）の場合、まず図１１より検索条件解析手段１１０にて、検索条件のタグ名を名称ＩＤに変換する（ステップ１１０１）。

【００７６】次にＣａｓｅ３と同様に、文字列索引検索手段１１１にて検索条件の文字列について検索処理を行ない、該当する検索単位識別子群を求める。最後に図１３より構造照合手段１１２にて、文字列索引検索手段１１１で求めた検索単位識別子群のうち、名称ＩＤがステップ１１０１で求めた名称ＩＤと一致するものだけを、要素管理テーブルを元に抽出し（ステップ１３０１）、最終的な検索結果とする。

【００７７】タグ名とその出現順序を指定した検索（Ｃａｓｅ２）の場合、Ｃａｓｅ１と同様な処理を行なった後、最後に出現順序照合処理（図１３のステップ１３０２）を行なう。ステップ１３０２では、該検索単位識別子のパス階層ＩＤをキーとしてパス階層インデックスを参照し、末端階層の出現順序が検索条件の出現順序と一致するものだけを抽出し、最終的な検索結果とする。

【００７８】パス名称とパス階層を指定した検索（Ｃａｓｅ４）の場合、検索条件解析手段１１０でＣａｓｅ３と同様にステップ１１０２の処理を行なった後、検索条件のパス階層をパス階層インデックスを用いてパス階層ＩＤへの変換を行なう（ステップ１１０３）。次にＣａｓｅ３と同様に、文字列索引検索手段１１１にて検索条件の文字列について検索処理を行ない、該当する検索単位識別子群を求める。

【００７９】最後に構造照合手段１１２にて、Ｃａｓｅ３と同様にパス名称ＩＤ照合処理（ステップ１３０３）を行なった後、パス階層ＩＤ照合処理（ステップ１３０４）を行なう。ステップ１３０４では、該検索単位識別子のパス階層ＩＤがステップ１１０３で変換したパス階層ＩＤと一致するものだけを抽出し、最終的な検索結果とする。

【００８０】最後に検索結果作成・表示処理について説明する。結果作成手段１１３は検索結果として得られた文書の書誌情報（タイトル、著者、日付など）を結果一覧表示用のデータとして、一覧データ格納手段１２１に格納する。このデータを結果表示手段１０４を通して端末１０１に表示する。次に端末１０１から実体表示要求としてこの検索結果一覧の中からどれか1つの文書が選択されると、結果作成手段１１１が実体データ格納手段１１５から指定された文書の実体を取得し、結果表示手段１０４を通して端末１０１に表示する。なお、構造解析手段１０７によって要素に分割された単位で、登録対象文書を実体データ格納手段１２０に登録しておくことにより、検索結果作成・表示処理において要素毎の結果一覧の作成、および要素毎の実体取得も可能である。

【００８１】以上のように本実施の形態では、構造化文書の論理構造情報を要素管理テーブル格納手段１１５、パス名称インデックス格納手段１１６、パス階層インデックス格納手段１１７、名称ＩＤテーブル格納手段１１８の４つに分けて格納し、文字列索引内部にこれら論理構造に関する情報を含めないことにより、文字列索引のサイズ縮小を可能とする。更に文書の特定の要素内容の追加、変更、削除を行なう際に、追加、変更、削除により論理構造の変化の発生した検索単位識別子のレコードについて、要素管理テーブルの変更処理を行なうだけで済むため、文字列索引内部に論理構造に関する情報を含める方法と比較して、処理量の大幅な軽減が可能となる。（文字列索引内部に論理構造に関する情報を含める方法の場合、追加、変更、削除により、論理構造の変化が発生した要素に関する全文字連鎖の文字列索引に対して修正処理が発生するため。）具体例を以下に示す。図１５は図３の構造をした文書の第１章第１節と第１章第２節の間に１５０１に示すノード群を追加した例である。この場合、１５０２のノードは第１章第２節から第１章第３節へと変更しなくてはならない。この時本実施の形態の方法では、既登録のデータに関しては、要素管理テーブルにおける検索単位識別子１０、および１１のレコードのパス名称ＩＤとパス階層ＩＤを変更するだけで済む。一方、文字列索引内部に論理構造に関する情報を含める方法の場合、検索単位識別子１０および１１の要素の全文字連鎖の文字列索引に対して論理構造情報の変更を行なわなくてはならない。（仮に、検索単位識別子１０の要素の内容が１００文字であったとすると、２文字連鎖で索引を作成している場合、９９個の文字連鎖の文字列索引に対して変更が必要となる）。

【００８２】また、本実施の形態では要素の論理構造位置を特定するためのＩＤをパス名称ＩＤとパス階層ＩＤの２つに分けているため、論理構造が複雑かつ膨大になった場合でも、公知例のように１種類のＩＤ（文脈識別子）で論理構造位置を特定する方法と比較して、ＩＤの総数を少なく押さえることが可能となる。

【００８３】なお、本実施の形態では１文書の構造化文書の登録、検索について説明したが、複数文書の場合でも同様の処理で実現が可能である。また本実施の形態で

は、一種類のＤＴＤにおけるパス名称ＩＤの作成方法について説明したが、本システムに複数の異なるＤＴＤの文書の登録要求が起こった場合においても、各ノードに個別なパス名称ＩＤを割当てることにより、論理構造を指定した検索が実現可能である。また、要素管理テーブル、パス名称インデックス、パス階層インデックス、名称ＩＤテーブルを一次記憶上に持つことにより、構造照合手段１１２の高速化が可能である。

【００８４】また本実施の形態は、構造化文書の管理を目的とする装置について説明を行ったが、必ずしも構造化文書に限らず、木構造で表現可能なデータを管理するために上述のパス名称インデックス及びパス階層インデックスを利用して実体要素（データの実体）を管理することも可能である。

【００８５】さらに実施の形態１は、装置として実現する例を示したが、その他に汎用計算機に本実施の形態に開示した構造化文書管理装置として機能するプログラムをインストールすることによっても実現することが可能である。

【００８６】（実施の形態２）以下、本発明の実施の形態２について説明する。図１６は実施の形態２における構造化文書管理装置の構成図である。実施の形態１の構成図である図１と異なるのは、データ格納部１０６にパス名称ＩＤ照合テーブル格納手段１６０１、パス階層ＩＤ照合テーブル格納手段１６０２を新たに備えているところである。またそれに伴い、検索条件解析手段１１０、および構造照合手段１１２の処理が実施の形態１とは異なる。

【００８７】パス名称ＩＤ照合テーブル格納手段１６０１は、各パス名称ＩＤが検索条件の構造指定の範囲内にあるかどうかの情報が格納される。

【００８８】パス階層ＩＤ照合テーブル格納手段１６０２は、各パス階層ＩＤが検索条件の構造指定の範囲内にあるかどうかの情報が格納される。

【００８９】実施の形態２における目的は、実施の形態１における検索条件の構造指定パターンＣａｓｅ１からＣａｓｅ４以外の構造指定に対応することである。Ｃａｓｅ１からＣａｓｅ４はタグ名やパス名称などで指定された末端要素そのものに対して検索を行なうものである。実施の形態２で実現する検索は、実体を持たない中間ノード以下を指定した検索である。例えば、「“章”以下に“管理”という文字列を含む文書を検索する」といった検索条件に対応することを目的とする。

【００９０】実施の形態２における登録処理は、実施の形態１と同様であるため説明を省略する。

【００９１】次に実施の形態２における検索処理の流れを具体例を示して説明する。なお、以下に示す本実施の形態における文書検索処理の説明においては、名称ＩＤテーブル、パス名称インデックス、パス階層インデックス、要素管理テーブルには、それぞれ図５、図６、図７、図８のようなデータが格納されているものとして説明を行なう。

【００９２】まず、検索条件入力手段１０３を通して、端末１０１から「パス名称が“／論文／本文／章”である中間ノード以下である要素に、“管理”という文字列が含まれる文書」という条件が与えられたとする。

【００９３】図１７は実施の形態２における検索条件解析手段１１０の処理の流れを示した図である。ここでの例では検索条件の構造指定としてパス名称以下が指定されているので、図１７のＣａｓｅ７に該当する。Ｃａｓｅ７ではステップ１１０２で、実施の形態１と同様に検索条件のパス名称をパス名称ＩＤに変換する。パス名称インデックスが図６の場合、検索条件のパス名称“／論文／本文／章”はパス名称ＩＤ“Ｎ６”に変換される。次にステップ１７０１でパス名称ＩＤ照合テーブルを作成する。図１８はここでの検索条件の例におけるパス名称ＩＤ照合テーブルの内容を示す図である。このパス名称ＩＤ照合テーブルは、検索要求ごとに作成し、パス名称インデックスの全パス名称ＩＤについて、検索条件で指定された範囲内のパス名称ＩＤと範囲外のパス名称ＩＤを即座に判断するために作成する。この例の場合、図６のパス名称インデックスよりパス名称ＩＤ“Ｎ６”以下にあるパス名称ＩＤ“Ｎ７、Ｎ８、Ｎ９、Ｎ１０、Ｎ１１”が範囲内で、それ以外は範囲外となる。

【００９４】次に文字列索引検索手段１１１で、検索条件の文字列について検索処理を行なう。処理手順は実施の形態１と同様であるため省略するが、ここでの例である“管理”という文字列で検索した結果として、検索単位識別子“１”と“９”が得られたものとして、説明を続ける。

【００９５】次に構造照合手段１１２で、文字列索引検索手段１１１で得られた検索単位識別子群の中から、検索条件の構造指定を満たす最終的な検索結果を求める。図１９は実施の形態２における構造照合手段１１２の処理の流れを示した図である。

【００９６】図１９におけるＣａｓｅ５からＣａｓｅ８というのは、図１７の検索条件の構造指定パターンＣａｓｅ５からＣａｓｅ８と同様である。ここでの例では、Ｃａｓｅ７（パス名称以下を指定）であるので、ステップ１３０３のパス名称ＩＤ照合処理を行なう。ただし、Ｃａｓｅ７におけるパス名称ＩＤ照合処理は、パス名称ＩＤ照合テーブルを用いて照合を行なう。図２０はこの例における構造照合処理の詳細を示す図である。まず文字列索引検索手段１１１で得られた検索単位識別子群（２００１）をキーとして要素管理テーブルを参照する。そこで該検索単位識別子のパス名称ＩＤからパス名称ＩＤ照合テーブルを参照し、照合フラグが“１”（範囲内）であるものだけを最終的な検索結果とする。

【００９７】なお、本実施の形態では、検索条件の構造指定として、タグ名で指定された中間ノード以下に対す

る検索（Ｃａｓｅ５）、タグ名とその出現順序で指定された中間ノード以下に対する検索（Ｃａｓｅ６）、パス名称とパス階層で指定された中間ノード以下に対する検索（Ｃａｓｅ８）にも対応可能である。以下でそれぞれＣａｓｅでの処理について簡潔に説明する。

【００９８】タグ名で指定された中間ノード以下に対する検索（Ｃａｓｅ５）の場合、検索条件解析手段１１０と文字列索引検索手段１１１における処理は、実施の形態１のＣａｓｅ１と同様であるため省略する。最後に図１９より構造照合手段１１２にて構造指定のチェックを行なう。ここでステップ１９０１のパス名称ＩＤ作成・更新・照合処理について説明する。図２１はパス名称ＩＤ作成・更新・照合処理の流れを示したフローチャートであり、このフローチャートに沿って説明する。

【００９９】まずパス名称ＩＤ照合テーブルの照合フラグを“０”（未定）で初期化しておく（ステップ３１０１）。次に文字列索引検索手段１１１で求めた検索単位識別子群それぞれについて以下の処理を繰り返す。まず検索単位識別子を取得し（ステップ３１０２）、該検索単位識別子のパス名称ＩＤ（要素管理テーブルより取得）の照合フラグを参照（ステップ３１０３）し、該照合フラグが“１”（範囲内）であれば（ステップ３１０４）、該検索単位識別子を最終的な検索結果に含める（ステップ３１０５）。照合フラグが“２”（範囲外）であれば（ステップ３１０６）、該検索単位識別子は最終的な検索範囲に含めない（ステップ３１０７）。照合フラグが“０”（未定）であったら、該検索単位識別子のパス名称ＩＤをキーとしてパス名称インデックスを参照し（ステップ３１０８）、検索条件解析手段１１０のステップ１１０１で求めた名称ＩＤと一致するか、もしくは、たどったノードのパス名称ＩＤの照合フラグが“１”（範囲内）の場合（ステップ３１０９）、該検索単位識別子のパス名称ＩＤと、そこまでたどったパス名称ＩＤ全てに対して、パス名称ＩＤ照合テーブルの照合フラグを１に設定し（ステップ３１１０）、該検索単位識別子を最終的な検索結果に含める。

【０１００】逆に、たどったノードのパス名称ＩＤの照合フラグが“２”（範囲外）の場合（ステップ３１１１）、該検索単位識別子のパス名称ＩＤと、そこまでたどったパス名称ＩＤ全てに対して、パス名称ＩＤ照合テーブルの照合フラグを“２”（範囲外）に設定し（ステップ３１１２）、該検索単位識別子を最終的な検索結果に含めない。

【０１０１】さらに、たどったノードのパス名称ＩＤの照合フラグが“０”（未定）の場合は、１階層登り（ステップ３１１３）、ルートノードであるか否かを判定し（ステップ３１１４）し、ルートノードでなければ、再びステップ３１０８に戻る。ルートノードである場合は、該検索単位識別子のパス名称ＩＤと、それまでたどったパス名称ＩＤ全ての照合フラグを２“範囲外”に設定する（ステップ３１１２）。

【０１０２】次の該当検索単位識別子が存在するか否かをチェックし（ステップ３１１５）、存在する場合は、ステップ３１０２へ戻る。存在しない場合は、本処理を終了する。

【０１０３】このように徐々に各パス名称ＩＤが検索条件の範囲内にあるかどうかのパス名称ＩＤ照合テーブルが学習されていくため、別の検索単位識別子に対してパス名称ＩＤの照合を行なう際に、すでに範囲内であると判明している（照合フラグが“１”である）パス名称ＩＤであった場合、該検索単位識別子を即座に最終的な検索結果に含ませることが可能となる。

【０１０４】なお上記ステップ３１０１からステップ３１１５までの処理については、汎用計算機に上記ステップの処理を実現するプログラムをインストールすることにより実現することが可能である。

【０１０５】また上記実施の形態では、構造化文書において中間ノード以下を検索範囲に指定した場合に、検索範囲に含まれるノードを決定する例を示したが、構造化文書に限らず、その他木構造で表現できるデータについても同様に適用することが可能である。

【０１０６】タグ名とその出現順序で指定された中間ノード以下に対する検索（Ｃａｓｅ６）の場合、検索条件解析手段１１０、文字列索引検索手段１１１、および構造照合手段１１２のステップ１９０１まではＣａｓｅ５と同様の処理を行なう。次にステップ１９０１でパス名称ＩＤが範囲内にあった場合に限り、ステップ１９０２のパス階層ＩＤ作成・更新・照合処理を行なう。図２２はパス階層ＩＤ照合テーブルの例である。ステップ１９０２ではステップ１９０１のパス名称ＩＤに関する処理と同様に、パス階層ＩＤについて構造指定の範囲にあるかどうか学習していき、照合フラグが“１”のパス階層ＩＤを持つ検索単位識別子を最終的な検索結果とする。

【０１０７】パス名称とパス階層で指定された中間ノード以下に対する検索（Ｃａｓｅ８）の場合、検索条件解析手段１１０では、Ｃａｓｅ７と同様な処理を行なったあとに、ステップ１７０２にてパス階層ＩＤ照合テーブルを作成する。このパス階層ＩＤ照合テーブルは、パス階層インデックスにおいて、ステップ１１０３で求めたパス階層ＩＤにあたるノードとそれ以下全てのノードのパス階層ＩＤに対する照合フラグを“１”（範囲内）に、それ以外を“２”（範囲外）に設定する。文字列索引検索手段１１１での処理はＣａｓｅ７と同様であるため説明を省略する。

【０１０８】次に構造照合手段１１２において、Ｃａｓｅ７と同様な処理を行なった後、ステップ１７０２にて作成したパス階層ＩＤ照合テーブルを用いて、該検索単位識別子のパス階層ＩＤの照合処理を行なう。ここでパス階層ＩＤ照合テーブルの照合フラグが“１”であるパス階層ＩＤを持つ検索単位識別子のみ、最終的な検索結

果とする。

【０１０９】実施の形態２における検索結果作成・表示処理は実施の形態１と同様であるため、説明を省略する。

【０１１０】以上のように本実施の形態では、中間ノードを以下を指定した検索の際に、各パス名称ＩＤが検索条件の構造指定の範囲内にあるかどうかの情報が格納されるパス名称ＩＤ照合テーブルや、各パス階層ＩＤが検索条件の構造指定の範囲内にあるかどうかの情報が格納されるパス階層ＩＤ照合テーブルを作成し、構造照合処理を行なうことにより、中間ノード以下を指定した高速な検索を実現する。

【０１１１】なお、図１６に示す実施の形態２の構成においても、パス名称ＩＤ照合テーブル格納手段１６０１、およびパス階層ＩＤ照合テーブル格納手段１６０２を使用しないことにより、実施の形態１における検索条件の構造指定Ｃａｓｅ１からＣａｓｅ４にも、対応可能である。また本実施の形態の説明において、パス名称ＩＤ照合テーブル、およびパス階層ＩＤ照合テーブルの照合フラグの値を、範囲内の場合“１”、範囲外の場合“２”、未定の場合“０”としていたが、この照合フラグの値は範囲内、範囲外、未定の状態を判別可能な値であればどのような値を割当てても構わない。

【０１１２】さらに実施の形態２は、装置として実現する例を示したが、その他に汎用計算機に本実施の形態に開示した構造化文書管理装置として機能するプログラムをインストールすることによっても実現することが可能である。

【０１１３】（実施の形態３）以下、本発明の実施の形態３について説明する。実施の形態３における構造化文書管理装置の構成図は実施の形態１における図１、もしくは実施の形態２における図１６と同様である。ただし、文字列索引作成手段１０９における文字列索引の作成方法が実施の形態１および実施の形態２とは若干異なり、それに伴い文字列索引検索手段１１１と構造照合手段１１２における処理が実施の形態１および実施の形態２とは異なる。

【０１１４】ここで実施の形態３における登録処理の流れについて説明する。まず構造化文書入力手段１０２、構造解析手段１０７、および構造情報作成手段１０８の処理は、実施の形態１および実施の形態２と同様であるため説明を省略する。

【０１１５】図２３は実施の形態３における文字列索引作成手段１０９の処理の流れである。ステップ９０１からステップ９０３までは実施の形態１および実施の形態２と同様であるため説明を省略する。次に該要素がＭｉｘｅｄ　Ｃｏｎｔｅｎｔを含むかどうか調べ（ステップ２２０１）、含む場合はこのＭｉｘｅｄ　Ｃｏｎｔｅｎｔに割当てられている検索単位識別子を取得する（ステップ２２０２）。この「Ｍｉｘｅｄ　Ｃｏｎｔｅｎｔ」とは、要素実体の内部で、該要素の子要素として存在する、要素実体のことである。例えば、図２４の２３１０のように、「段落」を表す要素の中で、更に「キーワード」タグに囲まれた要素がＭｉｘｅｄＣｏｎｔｅｎｔである。他の例としては、「強調」や「斜体」などがあり、検索する際には、これら「段落」と「キーワード」の要素にまたがった文字列でも検索してヒットすることが望まれる。そのためステップ２２０３で文字連鎖を取り出す際に、Ｍｉｘｅｄ　Ｃｏｎｔｅｎｔにまたがる文字連鎖も抽出し、ＭｉｘｅｄＣｏｎｔｅｎｔにまたがる文字連鎖の場合には、ステップ２２０４で文字列索引に、文字連鎖１文字目の検索単位識別子と文字連鎖２文字目の検索単位識別子と文字位置番号を格納する（以下、このようなＭｉｘｅｄ　Ｃｏｎｔｅｎｔにまたがる文字連鎖の文字列索引を、拡張文字列索引と呼ぶ）。この場合の文字位置番号は、該文字連鎖先頭文字がＭｉｘｅｄ　Ｃｏｎｔｅｎｔの外側の要素の中で何番目の文字かを表す番号とする。ステップ９０６から９０８までの処理は、実施の形態１および実施の形態２と同様であるため説明を省略する。

【０１１６】次にＭｉｘｅｄ　Ｃｏｎｔｅｎｔを含む要素の文字列索引の作成例について、図２４を用いて説明する。図２４の２３１０に示すように、「段落」の中に「キーワード」タグで囲まれたＭｉｘｅｄ　Ｃｏｎｔｅｎｔを含み、「キーワード」タグの要素の検索単位識別子は“１０１”、「段落」タグの要素の検索単位識別子は“１０２”が割当てられているものとして説明する。この例の場合に作成される文字列索引を図示したものが２３２０である。この例の場合、“を検”（２３２１）と“索す”（２３２３）の文字連鎖がＭｉｘｅｄ　Ｃｏｎｔｅｎｔにまたがっており、文字連鎖１文字目と文字連鎖２文字目の、２個の検索単位識別子が文字列索引に格納される。なお、図２４の２３２０は文字列索引の一部しか示されていないが、実際は登録対象文書の全要素の全文字列について文字列索引が作成される。

【０１１７】次に実施の形態３における文書検索の処理の流れについて説明する。まず検索条件入力手段１０３、検索条件解析手段１１０における処理は実施の形態１および実施の形態２と同様であるため説明を省略する。次に文字列索引検索手段１１１における処理についてだが、基本的には実施の形態１および実施の形態２と同様である。ただし実施の形態３では、文字列索引作成手段１０９において、Ｍｉｘｅｄ　Ｃｏｎｔｅｎｔにまたがる文字連鎖の場合、文字連鎖１文字目と文字連鎖２文字目の、２個の検索単位識別子含む拡張文字列索引を作成しているため、この拡張文字列索引が絡む場合の検索処理が新たに必要となる。以下、その具体例について図２４を用いて説明する。検索文字列が“検索する”である場合、２３１０の要素に該当する文字連鎖の文字列索引として２３２２、２３２３、２３２４が得られる。

ここで２３２２の検索単位識別子と、拡張文字列索引である２３２３の文字連鎖１文字目検索単位識別子が“１０１”で一致する。更に文字位置番号が“4”と“5”で連続している。また、拡張文字列索引２３２３の文字連鎖２文字目検索単位識別子と２３２４の検索単位識別子が“１０２”で一致し、更に文字位置番号が“5”と“6”で連続している。このような場合に文字連鎖２３３３から２３２４にかけて文字列検索にヒットしたことになる。その際、文字列検索結果の検索単位識別子として、検索文字列の先頭文字および末端文字に該当するの検索単位識別子のセットを返す。ここでの例の場合、先頭文字検索単位識別子“１０１”、末尾文字検索単位識別子“１０２”のセットを返す。次に構造照合手段の処理についてだが、基本的には実施の形態１および実施の形態２と同様である。ただし実施の形態３では、文字列索引検索手段１１１から得られる文字列検索結果群の中に、先頭文字検索単位識別子と末尾文字検索単位識別子のセットが含まれる場合があり、この場合の構造照合処理が新たに必要となる。

【０１１８】上記実施の形態３における文字列索引検索手段１１１の説明で用いた例では、文字列検索処理結果として、先頭文字検索単位識別子“１０１”、末尾文字検索単位識別子“１０２”のセットを返した。この場合、検索単位識別子“１０１”および“１０２”の両方に対して、実施の形態１および実施の形態２と同様な構造照合処理を行ない、両検索単位識別子とも検索条件の構造指定に当てはまる場合のみ、最終的な検索結果とする。

【０１１９】実施の形態３における検索結果作成・表示処理は実施の形態１および実施の形態２と同様であるため、説明を省略する。

【０１２０】以上のように本実施の形態では、登録対象構造化文書中にＭｉｘｅｄ　Ｃｏｎｔｅｎｔを含む場合に、Ｍｉｘｅｄ　Ｃｏｎｔｅｎｔにまたがる文字連鎖に対しても文字列索引（文字連鎖１文字目と文字連鎖２文字目の、２個の検索単位識別子を記憶する拡張文字列索引）を作成することによって、Ｍｉｘｅｄ　Ｃｏｎｔｅｎｔにまたがる文字列を検索対象とすることが可能となる。また、ＭｉｘｅｄＣｏｎｔｅｎｔである要素（上記説明では「キーワード」要素）を指定した検索も可能となる。

【０１２１】なお、実施の形態３の説明においては、２文字ずつ文字連鎖を取り出してそれぞれに文字列索引を作成しているが、この文字連鎖は２文字ずつでなくても構わない。この場合、実施の形態３における「文字連鎖１文字目検索単位識別子」を「文字連鎖先頭文字の検索単位識別子」に、「文字連鎖２文字目検索単位識別子」を「文字連鎖末尾文字の検索単位識別子」に置き換えることにより、同様の効果が実現可能である。

【０１２２】さらに実施の形態３は、装置として実現する例を示したが、その他に汎用計算機に本実施の形態に開示した構造化文書管理装置として機能するプログラムをインストールすることによっても実現することが可能である。

【０１２３】（実施の形態４）以下、本発明の実施の形態４について説明する。図２５は実施の形態４における構造化文書管理装置の構成図である。実施の形態１の構成図である図１と異なるのは、検索エンジン１０５に数値型索引作成手段２４０１と数値型索引検索手段２４０２を、データ格納部１０６に数値型設定格納手段２４０３と数値型索引格納手段２４０４を新たに備えているところである。

【０１２４】数値型索引作成手段２４０１は、あらかじめ設定されたタグ名の要素内容に対する数値範囲検索用の索引を作成する。

【０１２５】数値型索引検索手段２４０２は、数値型索引作成手段２４０１で作成された数値型索引を用いて数値範囲の検索処理を行なう。

【０１２６】数値型設定格納手段２４０３は、あらかじめ数値型索引を作成するように定められた要素のタグ名の集合が格納されている。

【０１２７】数値型索引格納手段２４０４は、数値型索引作成手段２４０１で作成された数値型索引を格納する。

【０１２８】ここで、実施の形態４における登録処理の流れについて具体例を用いて説明する。まず実施の形態４においては、本システムに文書を登録する前に、数値型設定格納手段２４０３にあらかじめ数値索引を作成する要素のタグ名として“価格”というタグ名が設定されているものとする。この時、図２６のような文書を登録する場合について説明する。構造化文書入力手段１０２、構造解析手段１０７、構造情報作成手段１０８、および文字列索引作成手段１０９の処理は、実施の形態１および実施の形態２と同様であるため説明を省略する。

【０１２９】図２７は実施の形態４における数値型索引作成手段２４０１の処理の流れである。まずステップ２６０１で登録文書の構造解析済みデータを読み込む。次に現在参照中の要素が数値型設定格納手段２４０３で数値型索引を作成するよう設定された要素かどうか調べ（ステップ２６０２）、設定されていない要素であったらステップ２６０６へ進む。設定されていた要素であったら、構造解析手段１０７のステップ４０６にて該要素に割当てられた検索単位識別子を取得する。次にステップ２６０４で該要素の実体（文字列）を数値データに変換する。その際、文字列が数字だけでなく単位などの文字データを含んでいる場合、数字部分の文字列だけ取り出し、数値データに変換する。そして数値型索引に該要素の検索単位識別子と数値データのレコードを追加する。この際、数値型索引は数値型設定格納手段２４０３で設定された要素のタグ名の名称ＩＤごとに作成する

（ステップ２６０５）。次にステップ２６０６で登録対象文書の全要素についてステップ２６０２から２６０５の処理を終了したか調べ、まだ未処理の要素が存在したらステップ２６０２以降の処理を繰り返す。全要素についてステップ２６０２から２６０５の処理を終了したら、最後にここで作成した数値型索引を数値型索引格納手段２４０４に追加する（ステップ２６０７）。

【０１３０】ここでの例の場合、数値型索引を作成する要素は図２６の２５０１に示す要素である。該要素の検索単位識別子が“２０１”であるとした場合に作成される数値型索引は図２８の２７１０のようになる。なお、図２８では数値データをLong型整数として格納しているが、Double型浮動小数点数などで格納することも可能である。ただし、名称ＩＤ単位で作成される数値型索引ごとに型を統一する必要がある。

【０１３１】次に実施の形態４における文書検索の処理について説明する。実施の形態４では、数値型設定格納手段２４０３で設定されたタグ名の要素に対して数値型索引を作成しているため、実施の形態１および実施の形態２で説明した構造を指定した文字列の検索のほかに、数値範囲の検索が可能となる。

【０１３２】例として、まず検索条件入力手段１０３を通して、端末１０１から「タグ名が“価格”である要素の内容が“１５００円～１７００円”である文書」という条件が与えられたとする。この時検索条件解析手段１１０の処理は実施の形態１のＣａｓｅ１と同様であるため説明を省略する。

【０１３３】次に検索条件が数値範囲を指定した検索なので、文字列索引検索手段１１１ではなく、数値型索引検索手段２４０２の処理を行なう。ここでの例の場合、“価格”タグの名称ＩＤについて作成された数値型索引に図２８の２７２０のようなデータが格納されているとすると、１５００以上、１７００以下の数値データを持つものとして２７２１（検索単位識別子：５４）、２７２２（検索単位識別子：２０１）、２７２３（検索単位識別子：５４５）の３つを抽出する。

【０１３４】次に構造照合手段１１２にて、数値型索引検索手段２４０２の処理で抽出した検索単位識別子について、検索条件の構造指定チェックを行なう。ここでの例における構造照合手段１１２の処理は実施の形態１と同様であるため説明を省略する。なお、実施の形態４では数値範囲検索における構造指定として、上記実施の形態１におけるＣａｓｅ１のみでなく、Ｃａｓｅ２、Ｃａｓｅ３、Ｃａｓｅ４に対応可能である。それぞれのＣａｓｅにおける検索条件解析手段１１０および構造照合手段１１２における処理は、実施の形態１と同様であるため説明を省略する。

【０１３５】実施の形態４における検索結果作成・表示処理は実施の形態１と同様であるため、説明を省略する。

【０１３６】以上のように本実施の形態では、あらかじめ数値型設定格納手段２４０３で設定されたタグ名の要素に対して数値型索引作成手段２４０１にて数値型索引を作成することにより、要素内容を数値データとして扱った数値範囲の検索が可能となる。

【０１３７】なお実施の形態４における数値型索引は、図２８の２７２０のような構造であるとして説明したが、この数値型索引は指定された数値範囲に該当する検索単位識別子を抽出できるものであればどのような構造でも構わない。また、実施の形態４において、文字列索引作成手段１０９での処理の後に、数値型索引作成手段２４０１を行なうものとして説明したが、文字列索引作成手段１０９の処理手順である図４のステップ４０５にて、要素実体に出会った場合に、ステップ４０６と４０７の処理と平行して、数値型索引作成手段２４０１の処理手順である図２７のステップ２６０２からステップ２６０５の処理を行なうことも可能である。

【０１３８】さらに実施の形態４は、装置として実現する例を示したが、その他に汎用計算機に本実施の形態に開示した構造化文書管理装置として機能するプログラムをインストールすることによっても実現することが可能である。

【０１３９】（実施の形態５）以下、本発明の実施の形態５について説明する。図２９は実施の形態５における構造化文書管理装置の構成図である。

【０１４０】本実施の形態は、ネットワーク上に構造化文書管理装置の各機能が分散していることを特徴とするものである。

【０１４１】構造化文書登録部３００１は、構造化文書を読み込み、解析し、構造化文書の木構造を生成する機能を有している。文字列索引作成部３００２は、構造化文書登録部３００１で解析された構造化文書について、検索用索引を生成する機能を有している。文字列検索部３００３は、検索条件を読み込み、検索条件に該当する文字列を有している要素実体を検索する機能を有している。結果表示部３００４は、前記文字列検索部３００３で得られた検索結果を端末１０１に表示する機能を有している。なお、端末１０１及びデータ格納部１０６は実施の形態１で記載した機能と同一の機能を有しており、データ格納部１０６は上記各機能ブロックが作成した解析済構造化文書、文字列索引、検索結果等をネットワーク経由で受け取り記憶する。端末１０１は、使用者の指定した検索条件をネットワーク経由で文字列検索部３００３に送る。また、結果表示部３００４に記憶されている検索結果をネットワーク経由で受け取り、表示する機能を有している。以下、各機能ブロック毎に説明する。

【０１４２】構造化文書登録部３００１は、構造化文書入力手段１０２と構造解析手段１０７と構造情報作成手段１０８より構成されており、これら３つの手段は、実施の形態１で記載している機能と同一の機能を有してい

る。ただし、構造情報作成手段１０８で作成される要素管理テーブルは実施の形態１で記載した図８の形式の他、図３１または３２のように検索単位識別子とパス名称ＩＤ及びパス階層ＩＤとの対応関係を示した形式、または検索単位識別子と名称ＩＤとの対応関係を示した形式でも構わない。

【０１４３】なお、上記構造化文書登録部３００１の機能と同一の機能はプログラム形式で実行可能であり、このプログラムを記録した可搬型媒体を用いて汎用計算機にインストールすることにより、構造化文書登録部３００１と同一の機能を実現できる。

【０１４４】また上記構造化文書登録部３００１は、それ自体で装置としての機能も果たすことが可能である。

【０１４５】文字列索引作成部３００２は、文字列索引作成手段１０９と、数値型索引作成手段２４０１から構成されている。文字列索引作成手段１０９は実施の形態１に記載した機能と同一の機能を有している。数値型索引作成手段２４０１は実施の形態４に記載した機能と同一の機能を有している。ただし、数値型索引作成手段２４０１は、検索条件として特定の数値範囲に該当する文字列を検索する場合に必要となる構成要素であり、検索条件に数値範囲が含まれない場合は、数値型索引作成手段２４０１は不要である。

【０１４６】なお、上記文字列索引作成部３００２の機能と同一の機能はプログラム形式で実行可能であり、このプログラムを記録した可搬型媒体を用いて汎用計算機にインストールすることにより、文字列索引作成部３００２と同一の機能を実現できる。

【０１４７】また文字列索引作成部３００２は、それ自体で装置としての機能も果たすことが可能である。

【０１４８】文字列検索部３００３は、検索条件入力手段１０３と、検索条件解析手段１１０と、文字列索引検索手段１１１と、数値型索引検索手段２４０２と、構造照合手段１１２から構成されている。検索条件入力手段１０３、検索条件解析手段１１０と、文字列索引検索手段１１１と、構造照合手段１１２は、実施の形態１に記載の機能と同一の機能を有する。ただし、構造情報作成手段１０８で作成される要素管理テーブルが図３１の形式の場合は、検索条件としてタグ名を指定することはできず、パス名称またはパス階層を指定することができる。一方、要素管理テーブルが図３２の形式の場合は、検索条件としてタグ名のみを指定することができる。

【０１４９】なお、上記文字列検索部３００３の機能と同一の機能はプログラム形式で実行可能であり、このプログラムを記録した可搬型媒体を用いて汎用計算機にインストールすることにより、文字列索引部３４０３と同一の機能を実現できる。

【０１５０】また文字列検索部３００３は、それ自体で装置としての機能も果たすことが可能である。

【０１５１】数値型索引検索手段２４０２は実施の形態４に記載の機能と同一の機能を有する。ただし、数値型索引検索手段２４０２は、検索条件として特定の数値範囲に該当する文字列を検索する場合に必要となる構成要素であり、検索条件に数値範囲が含まれない場合は、数値型索引検索手段２４０２は不要である。

【０１５２】なお、上記数値型索引検索手段２４０２の機能と同一の機能はプログラム形式で実行可能であり、このプログラムを記録した可搬型媒体を用いて汎用計算機にインストールすることにより、数値型索引検索手段２４０２と同一の機能を実現できる。

【０１５３】図３０は、文字列検索部３００３の処理の流れを示したフローチャートである。

【０１５４】まず、使用者の指定した検索条件を読み込み（ステップ３００５）、次に、読み込んだ検索条件に該当する名称ＩＤ又は、パス名称ＩＤ又は、パス階層ＩＤのいずれかのＩＤ（以下ＩＤ１）に変換する（ステップ３００６）。なお、前記３つのＩＤのうち、いずれのＩＤに変換されるかは図１１に示すように使用者の検索条件に依存する。また、どのような検索条件が可能であるかは図８、図３１、図３２に示した要素管理テーブルの形式に制約される。次に、前記検索条件に該当する文字列を有するすべての検索単位識別子（以下、ＩＤ２）を特定する（ステップ３００７）し、前記ＩＤ２に基づいて要素管理テーブルを参照し、対応する名称ＩＤ又は、パス名称ＩＤ又は、パス階層ＩＤのいずれかのＩＤ（以下、ＩＤ３）を特定し（ステップ３００８）、最後に、前記ＩＤ１とＩＤ３が一致する検索単位識別子を特定する（ステップ３００９）。

【０１５５】結果表示部３００４は、結果作成手段１１３と結果表示手段１０４から構成されている。結果作成手段１１３と結果表示手段１０４は、実施の形態１に記載の機能と同一である。

【０１５６】さらに実施の形態５は、装置として実現する例を示したが、その他に汎用計算機に本実施の形態に開示した構造化文書管理装置として機能するプログラムをインストールすることによっても実現することが可能である。

【０１５７】

【発明の効果】以上のように、本発明によれば構造化文書の様々な論理構造を指定した検索が可能な構造化文書管理装置において、文字列索引内部に論理構造に関する情報を含めないことにより、文字列索引のサイズ縮小を可能とする効果を有する。更に文書の特定の要素内容の追加、変更、削除を行なう際に、処理量が大幅に軽減されるという効果を有する。

【０１５８】また、ノードの論理構造位置を特定するためのＩＤをパス名称ＩＤとパス階層ＩＤの２つに分けて管理しているため、論理構造が複雑かつ膨大になった場合でも、構造を特定するためのＩＤの総数を少なく押さえることを可能とする効果を有する。

【０１５９】また、各パス名称ＩＤが検索条件の構造指定の範囲内にあるかどうかの情報が格納されるパス名称ＩＤ照合テーブルや、各パス階層ＩＤが検索条件の構造指定の範囲内にあるかどうかの情報が格納されるパス階層ＩＤ照合テーブルを作成し、構造照合処理を行なうことにより、中間ノード以下を指定した高速な検索を実現するという効果を有する。

【０１６０】なお、上述したように従来の技術では検索範囲として中間ノード以下を指定した場合、たとえ同一の親ノードを持つ同一タグ名を有するノードでも異なる文脈識別子が割り当てられるため、検索条件に該当するか否かをチェックする為のＯＲ検索が必要となり、検索時間が大きくなるという課題を有していたが、本発明は、同一の親ノードを持つ同一タグ名を有するノードがたとえ複数存在しても、同一の識別子を付与するために、ＯＲ検索が不要となり、検索時間が短縮できるという効果を有する。

【０１６１】また、Ｍｉｘｅｄ　Ｃｏｎｔｅｎｔにまたがる文字連鎖に対して拡張文字列索引を作成することによって、Ｍｉｘｅｄ　Ｃｏｎｔｅｎｔにまたがる文字列を検索対象とすること、およびＭｉｘｅｄ　Ｃｏｎｔｅｎｔである要素を指定した検索を可能とする効果を有する。

【０１６２】また、あらがじめ設定されたタグ名の要素に対して数値型索引を作成することにより、要素内容を数値データとして扱った数値範囲の検索を可能とする効果を有する。

【図面の簡単な説明】

【図１】本発明の実施の形態１における構造化文書管理装置の構成図

【図２】本発明の実施の形態１における構造化文書の一例を示す図

【図３】本発明の実施の形態１における構造を解析した結果の木構造の一例を示す図

【図４】本発明の実施の形態１における構造情報作成手段の処理手順を示す図

【図５】本発明の実施の形態１における名称ＩＤを割当てた例を示す図

【図６】本発明の実施の形態１におけるパス名称インデックスの一例を示す図

【図７】本発明の実施の形態１におけるパス階層インデックスの一例を示す図

【図８】本発明の実施の形態１における要素管理テーブルの一例を示す図

【図９】本発明の実施の形態１における文字列索引作成手段の処理手順を示す図

【図１０】本発明の実施の形態１における文字列索引の一例を示す図

【図１１】本発明の実施の形態１における検索条件解析手段の処理手順を示す図

【図１２】本発明の実施の形態１における文字列索引を用いた検索処理の詳細を示す図

【図１３】本発明の実施の形態１における構造照合手段の処理手順を示す図

【図１４】本発明の実施の形態１における構造照合処理の詳細を示す図

【図１５】本発明の実施の形態１におけるノード群を追加した木構造の一例を示す図

【図１６】本発明の実施の形態２における構造化文書管理装置の構成図

【図１７】本発明の実施の形態２における構造条件解析手段の処理手順を示す図

【図１８】本発明の実施の形態２におけるパス名称ＩＤ照合テーブルの一例を示す図

【図１９】本発明の実施の形態２における構造照合手段の処理手順を示す図

【図２０】本発明の実施の形態２における構造照合処理の詳細を示す図

【図２１】本発明の実施の形態２における構造照合手段で、中間ノードを指定した場合の検索範囲に該当するノードを特定するための処理手順を示す図

【図２２】本発明の実施の形態２におけるパス階層ＩＤ照合テーブルの一例を示す図

【図２３】本発明の実施の形態３における文字列索引作成手段の処理手順を示す図

【図２４】本発明の実施の形態３における拡張文字列索引の一例を示す図

【図２５】本発明の実施の形態４における構造化文書管理装置の構成図

【図２６】本発明の実施の形態４における構造化文書の一例を示す図

【図２７】本発明の実施の形態４における数値型索引作成手段の処理手順を示す図

【図２８】本発明の実施の形態４における数値型索引の一例を示す図

【図２９】本発明の実施の形態５における構造化文書管理装置の構成図

【図３０】本発明の実施の形態５における文字列検索部の処理手順を示す図

【図３１】本発明の実施の形態５における要素管理テーブルの一例を示す図

【図３２】本発明の実施の形態５における要素管理テーブルの一例を示す図

【図３３】従来の技術における文書登録システムの構成を示す図

【図３４】従来の技術における構造インデックスの生成過程を示す図

【図３５】従来の技術における文字列インデックスの例を示した図

【図３６】従来の技術における構造インデックスの更新

方法を示した図

【符号の説明】

１０１…端末
１０２…構造化文書入力手段
１０３…検索条件入力手段
１０４…結果表示手段
１０５…検索エンジン
１０６…データ格納部
１０７…構造解析手段
１０８…構造情報作成手段
１０９…文字列索引作成手段
１１０…検索条件解析手段
１１１…文字列索引検索手段
１１２…構造照合手段
１１３…結果作成手段
１１４…構造解析済みデータ格納手段
１１５…要素管理テーブル格納手段
１１６…パス名称インデックス格納手段
１１７…パス階層インデックス格納手段
１１８…名称ＩＤテーブル格納手段
１１９…文字列索引格納手段
１２０…実体データ格納手段
１２１…一覧データ格納手段
１６０１…パス名称ＩＤ照合テーブル格納手段
１６０２…パス階層ＩＤ照合テーブル格納手段
２４０１…数値型索引作成手段
２４０２…数値型索引検索手段
２４０３…数値型設定格納手段
２４０４…数値型索引格納手段
３００１…構造化文書登録部
３００２…文字列索引作成部
３００３…文字列検索部
３００４…結果表示部

【図２】

```
<論文>
 <書誌>
  <タイトル>構造化文書管理</タイトル>
  <著者> ・・・ </著者 >
  <著者> ・・・ </著者 >
  <日付> ・・・ </日付>
 </書誌>
 <本文>
  <章 >
   <章タイトル>登録方法</章タイトル>
   <段落>登録方法においては、まず. . . </段落>
   <節>
    <節タイトル>文書入力手段</節タイトル>
    <段落>まず、構造化文書の入力は. . . </段落>
    <段落>実体管理においては、まず. . . </段落>
   </節>
   <節>
    <節タイトル>文書格納手段</節タイトル>
    <段落>文書格納の際には. . . </段落>
   </節>
  </章>
 </本文>
</論文>
```

【図５】

名称ID	名称
T1	論文
T2	書誌
T3	タイトル
T4	著者
T5	日付
T6	本文
T7	章
T8	章タイトル
T9	段落
T10	節
T11	節タイトル
⋮	⋮

名称IDテーブル

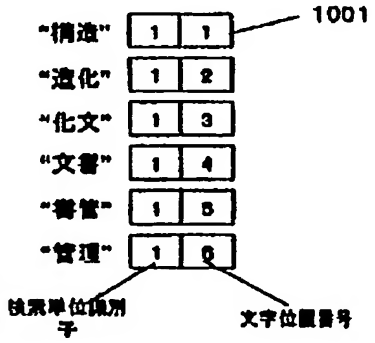
【図３２】

検索単位識別子	文書番号	名称ID
1	1	T3
2	1	T4
3	1	T4
4	1	T5
5	1	T8
6	1	T9
7	1	T11
8	1	T9
9	1	T9
10	1	T11
11	1	T9
⋮	⋯	⋯

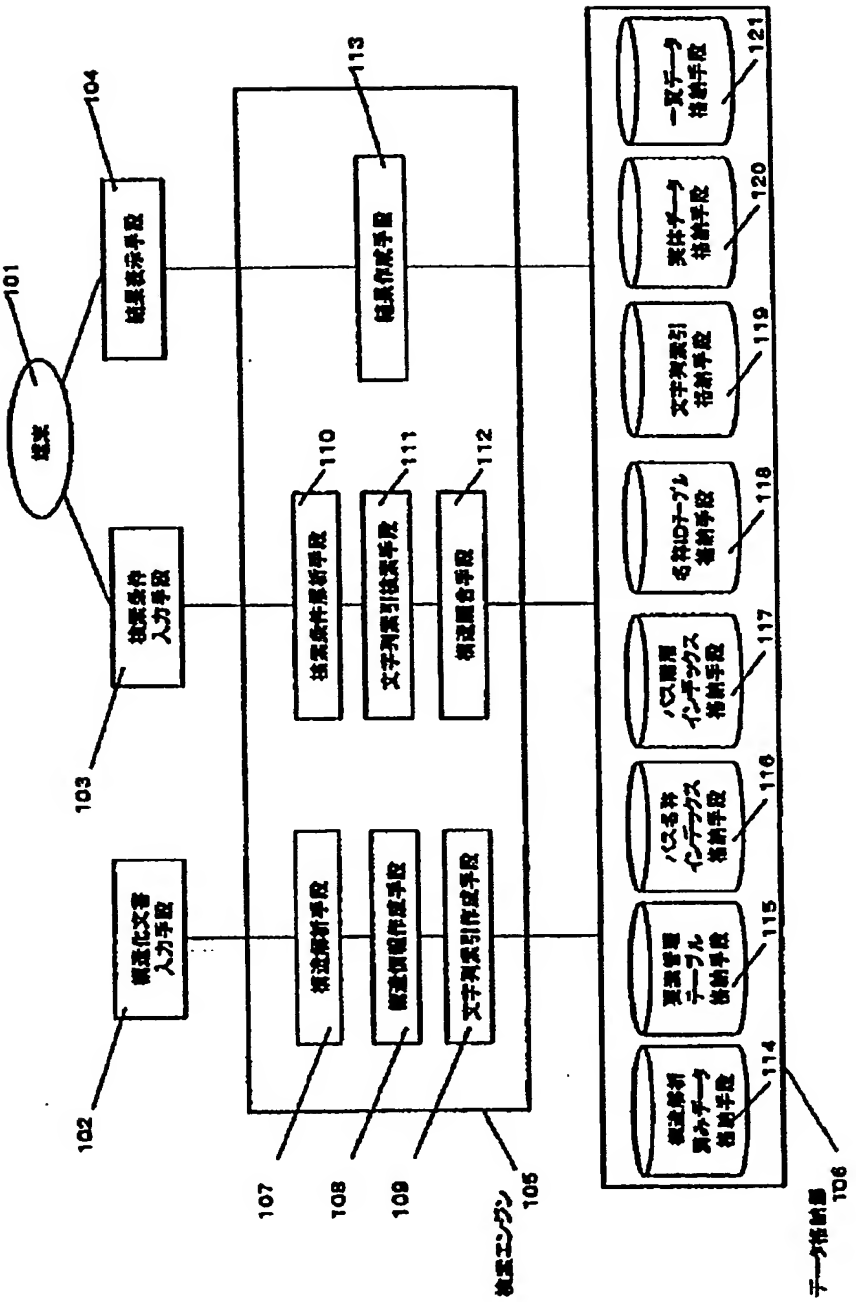
実体管理テーブル

【図１０】

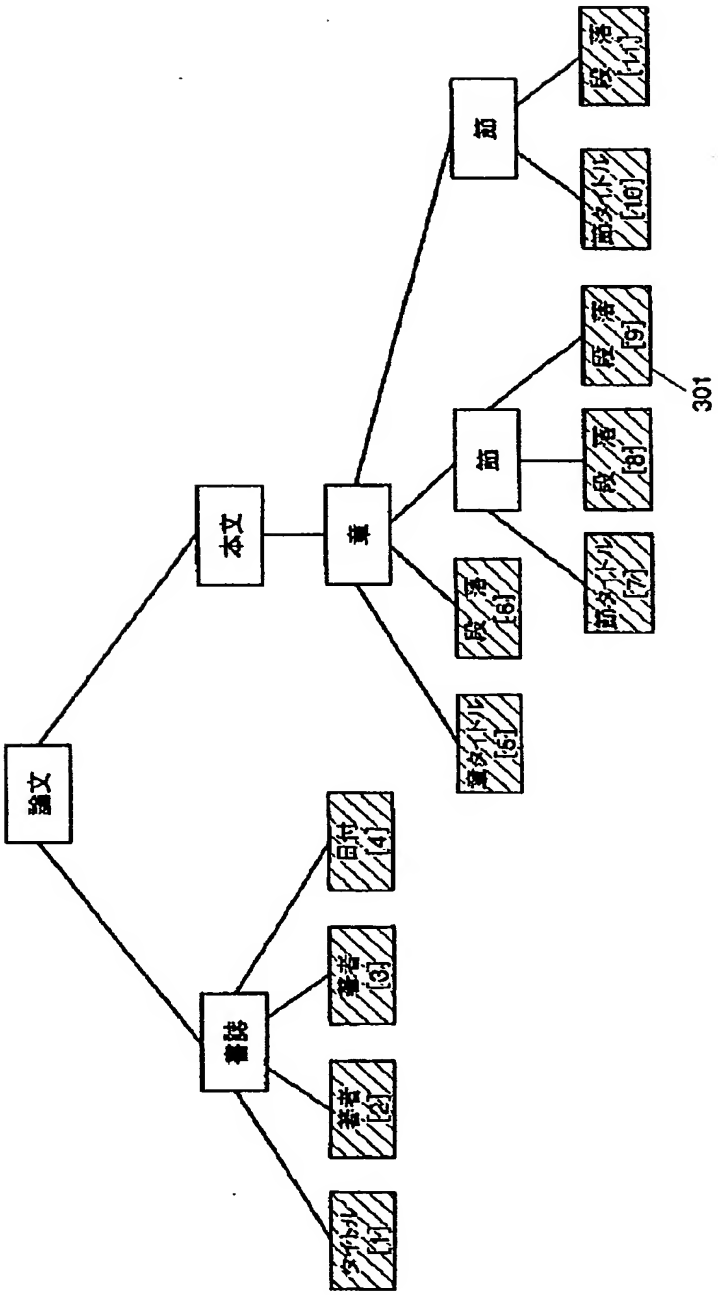
<タイトル> 構造化文書管理 </タイトル>



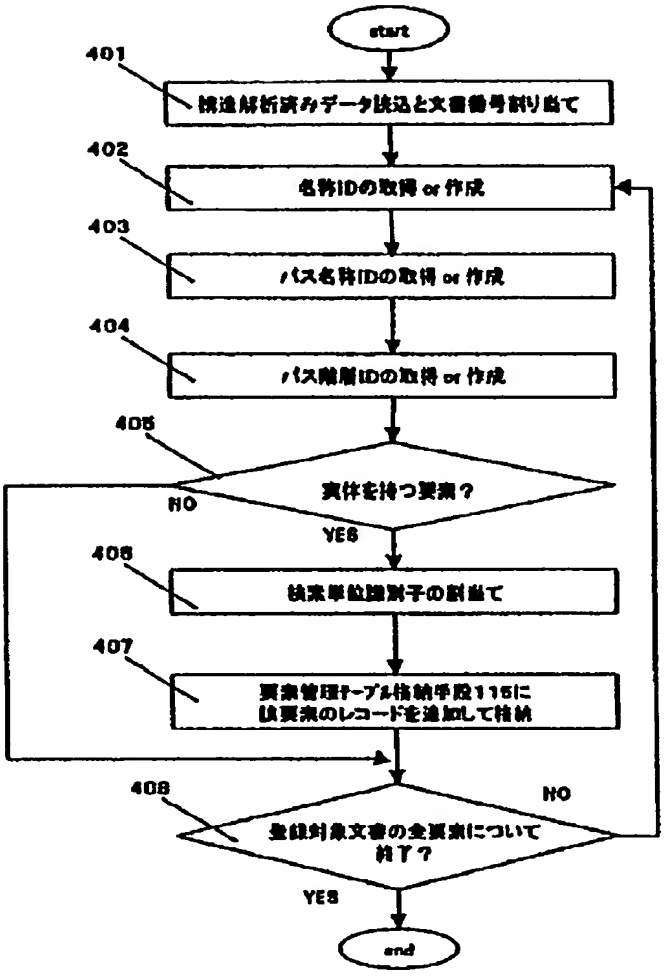
【図１】


101
端末
102
構造化文書入力手段
103
検索条件入力手段
104
結果表示手段
105
検索エンジン
106
データ格納部
107
構造解析手段
108
構造情報作成手段
109
文字列索引作成手段
110
検索条件展開手段
111
文字列索引検索手段
112
結果統合手段
113
結果作成手段
114
構造解析実体データ格納手段
115
異表管理テーブル格納手段
116
パス名称インデックス格納手段
117
パス順序インデックス格納手段
118
名称IDテーブル格納手段
119
文字列索引格納手段
120
実体データ格納手段
121
一文データ格納手段

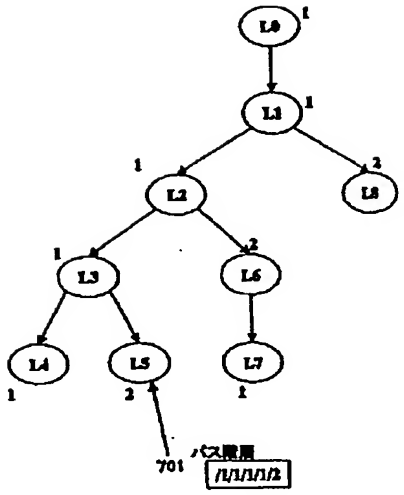
【図３】


論文
書誌
タイトル [1]
著者 [2]
著者 [3]
抄録 [4]
本文
章
段落 [6]
節タイトル [5]
節
段落 [7]
節タイトル [8]
節
段落 [9]
節タイトル [10]
段落 [11]
301

【図４】



【図７】



【図８】

検索単位識別子	文書番号	パス名称ID	パス階層ID	名称ID
1	1	N2	L2	T3
2	1	N3	L2	T4
3	1	N3	L8	T4
4	1	N4	L2	T5
5	1	N7	L3	T8
6	1	N8	L3	T9
7	1	N10	L4	T11
8	1	N11	L4	T9
9	1	N11	L5	T9
10	1	N10	L7	T11
11	1	N11	L7	T9
...	...	...	...	...

801

要素管理テーブル

【図１８】

パス名称ID	照合フラグ
N0	2
N1	2
N2	2
N3	2
N4	2
N5	2
N6	1
N7	1
N8	1
N9	1
N10	1
N11	1

0:未定
1:範囲内
2:範囲外

パス名称ID照合テーブル

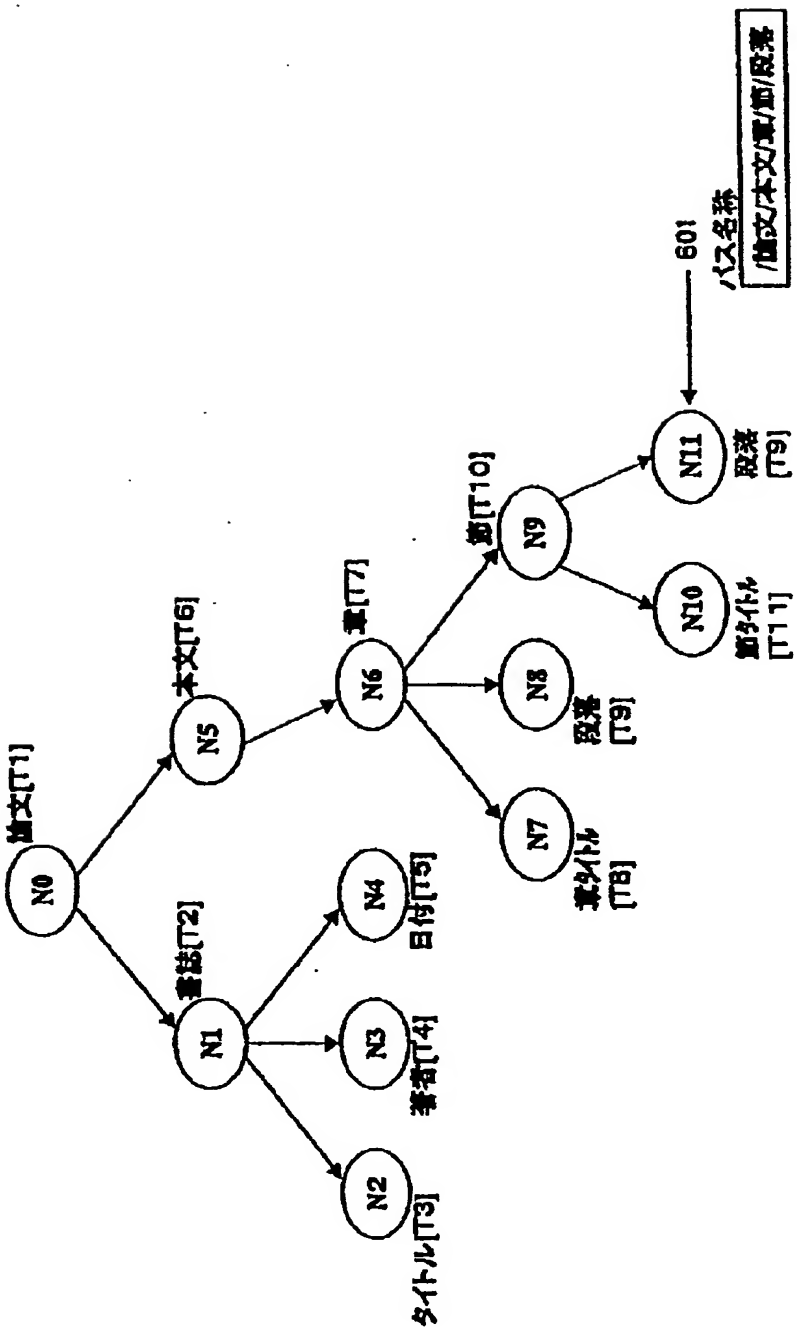
【図２２】

パス階層ID	照合フラグ
L0	2
L1	2
L2	1
L3	1
L4	1
L5	1
L6	1
L7	1
L8	2

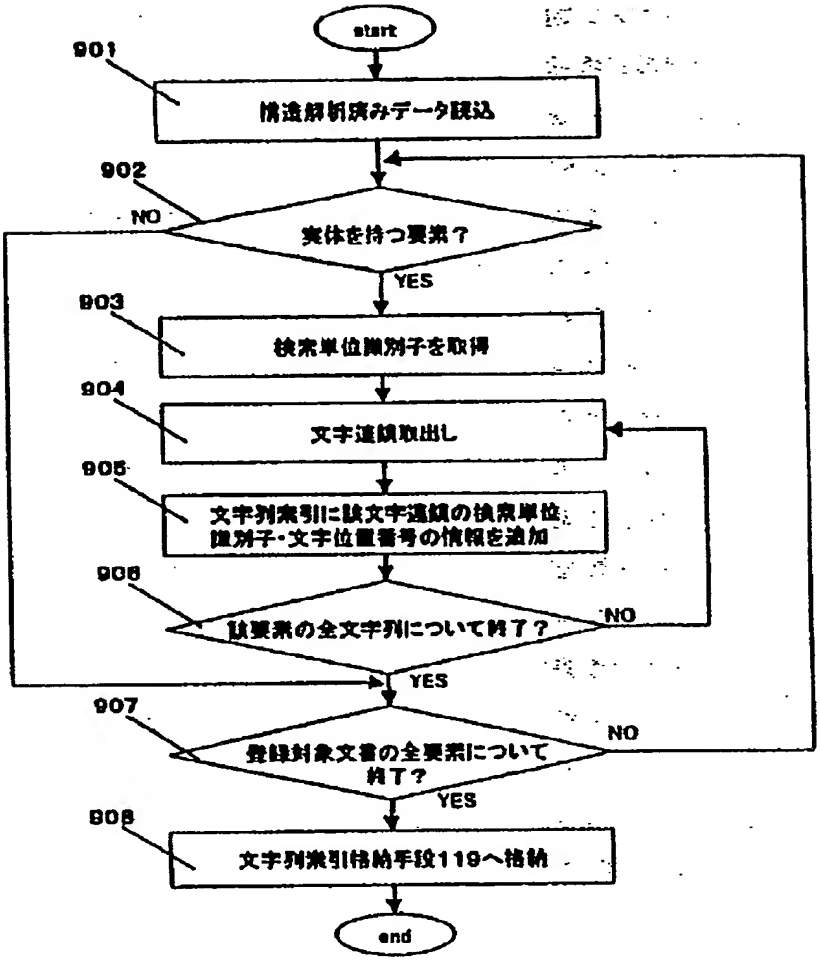
0:未定
1:範囲内
2:範囲外

パス階層ID照合テーブル

【図６】


N0
論文[T1]
N1
著者[T2]
N2
タイトル[T3]
N3
著者[T4]
N4
日付[T5]
N5
本文[T6]
N6
章[T7]
N7
章タイトル
[T8]
N8
段落
[T9]
N9
節[T10]
N10
節タイトル
[T11]
N11
段落
[T9]
601
パス名称
/論文/本文/章/節/段落

【図９】



【図２８】

2710

検索単位識別子	数値データ
201	000001600

2720

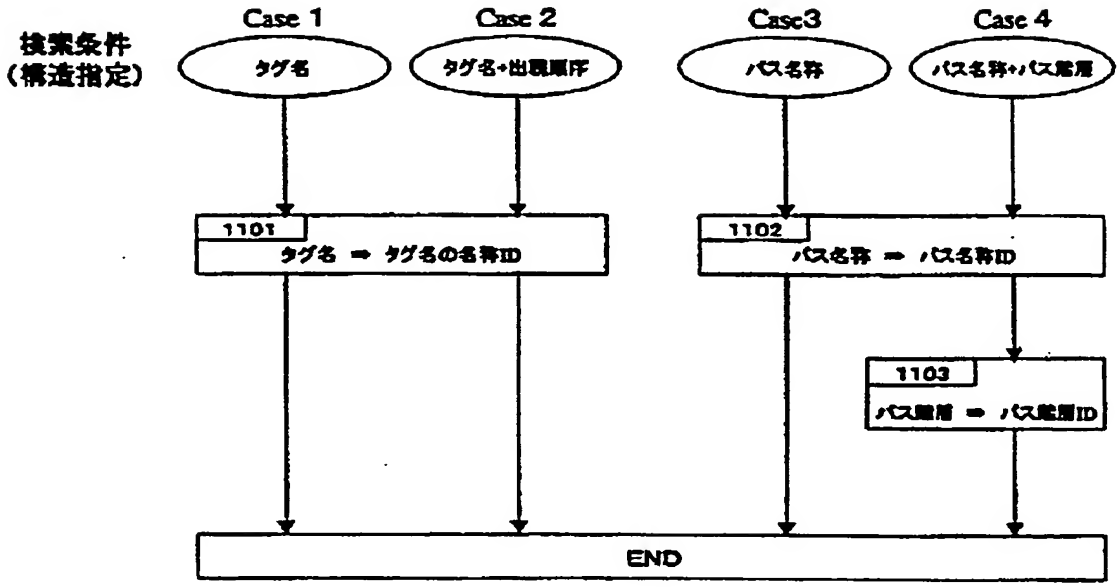
検索単位識別子	数値データ	
25	000002998	
54	000001596	2721
105	000019800	
201	000001600	2722
231	000000670	
545	000001540	2723
1352	000000800	

【図３１】

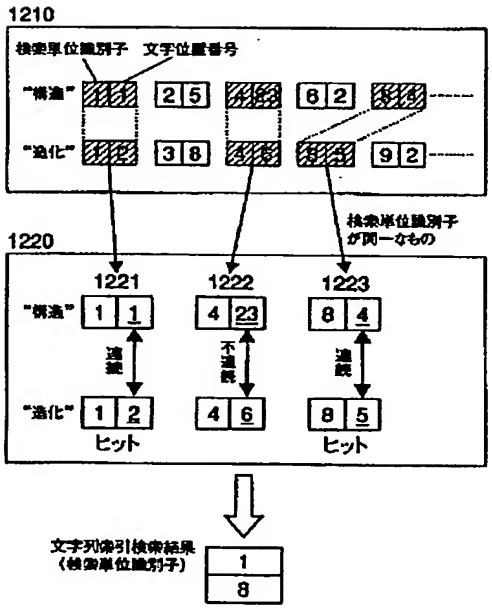
検索単位識別子	文書番号	パス名称ID	パス階層ID
1	1	N2	L2
2	1	N3	L2
3	1	N3	L8
4	1	N4	L2
5	1	N7	L3
6	1	N8	L3
7	1	N10	L4
8	1	N11	L4
9	1	N11	L5
10	1	N10	L7
11	1	N11	L7
...	...	...	...

要素管理テーブル

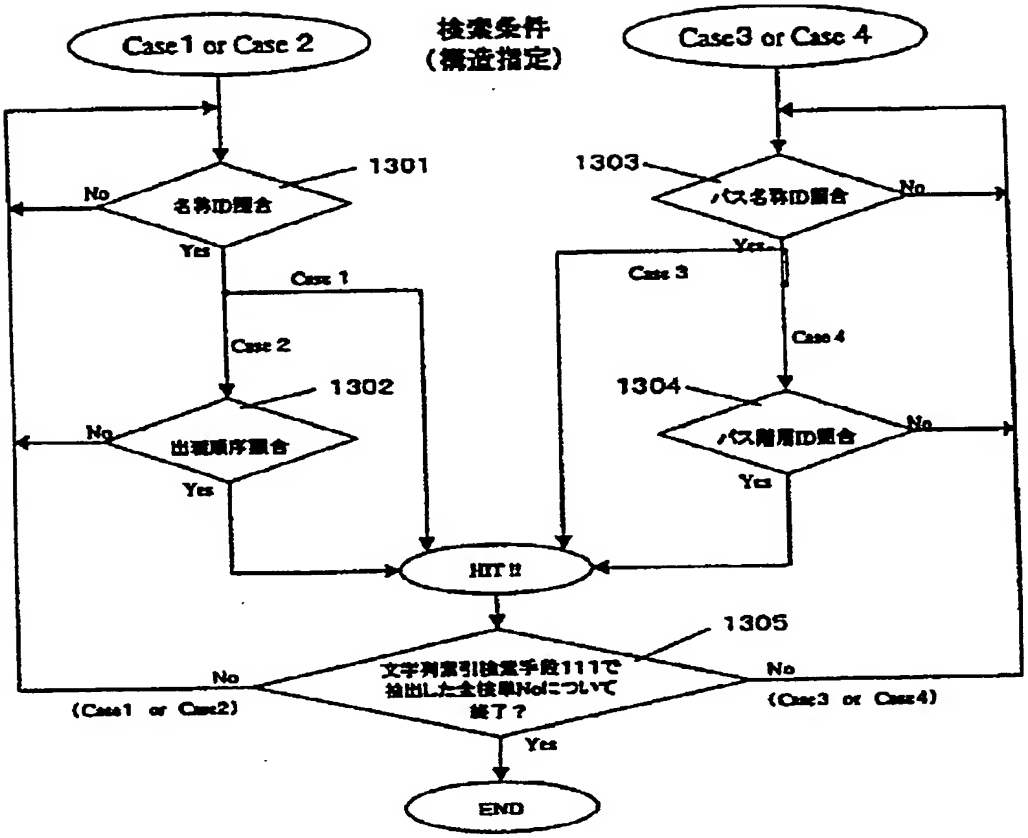
【図１１】


検索条件
(構造指定)
Case 1
タグ名
Case 2
タグ名+出現順序
Case3
パス名称
Case 4
パス名称+パス階層
1101
タグ名 ⇒ タグ名の名称ID
1102
パス名称 ⇒ パス名称ID
1103
パス階層 ⇒ パス階層ID
END

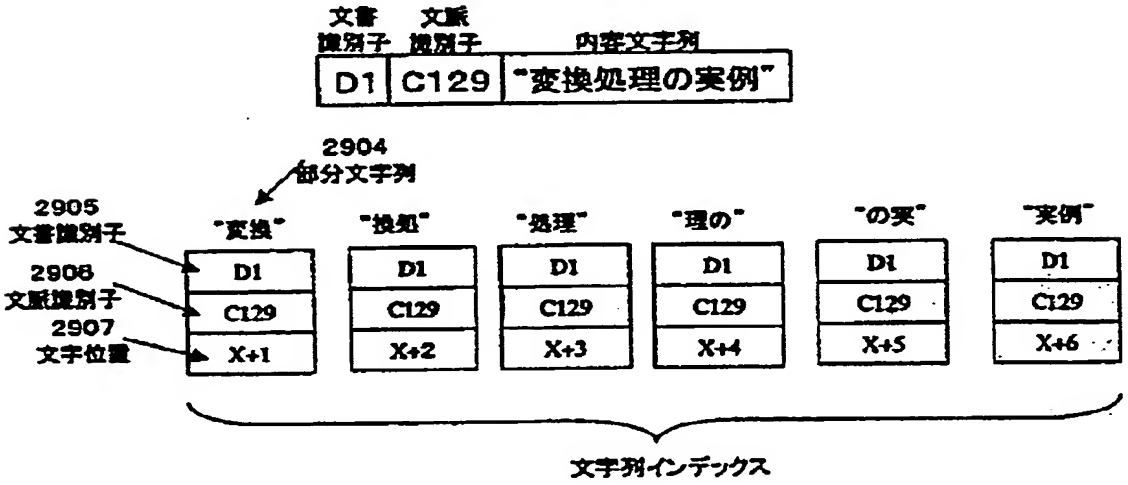
【図１２】


1210
検索単位識別子 文字位置番号
"構造"
2 5
6 2
"造化"
3 8
9 2
検索単位識別子
が同一なもの
1220
1221
1222
1223
"構造"
1 1
4 23
8 4
連続
不連続
連続
"造化"
1 2
4 6
8 5
ヒット
ヒット
文字列索引検索結果
(検索単位識別子)
1
8

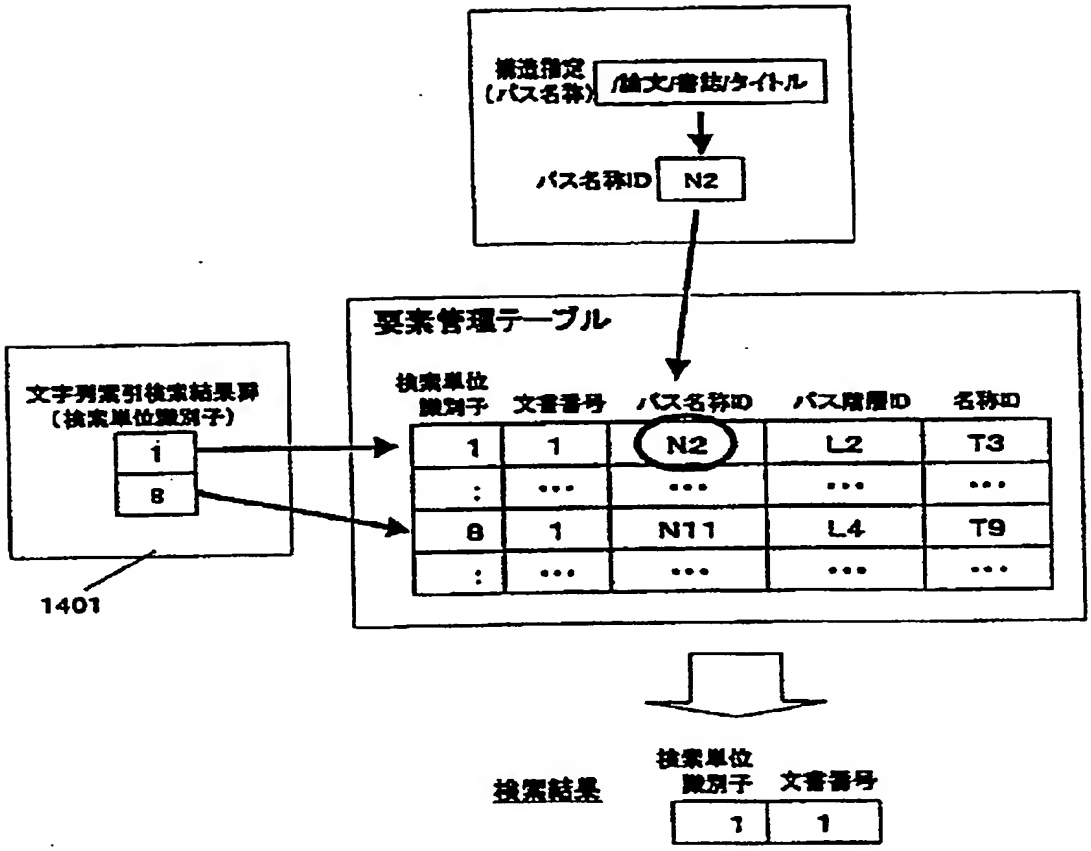
【図１３】


Case1 or Case 2
検索条件
(構造指定)
Case3 or Case 4
1301
名称ID照合
No
Yes
Case 1
Case 2
1302
出現順序照合
No
Yes
1303
パス名称ID照合
No
Yes
Case 3
Case 4
1304
パス階層ID照合
No
Yes
HIT !!
1305
文字列索引検索手段111で
抽出した全候補Noについて
終了？
No
(Case1 or Case2)
No
(Case3 or Case4)
Yes
END

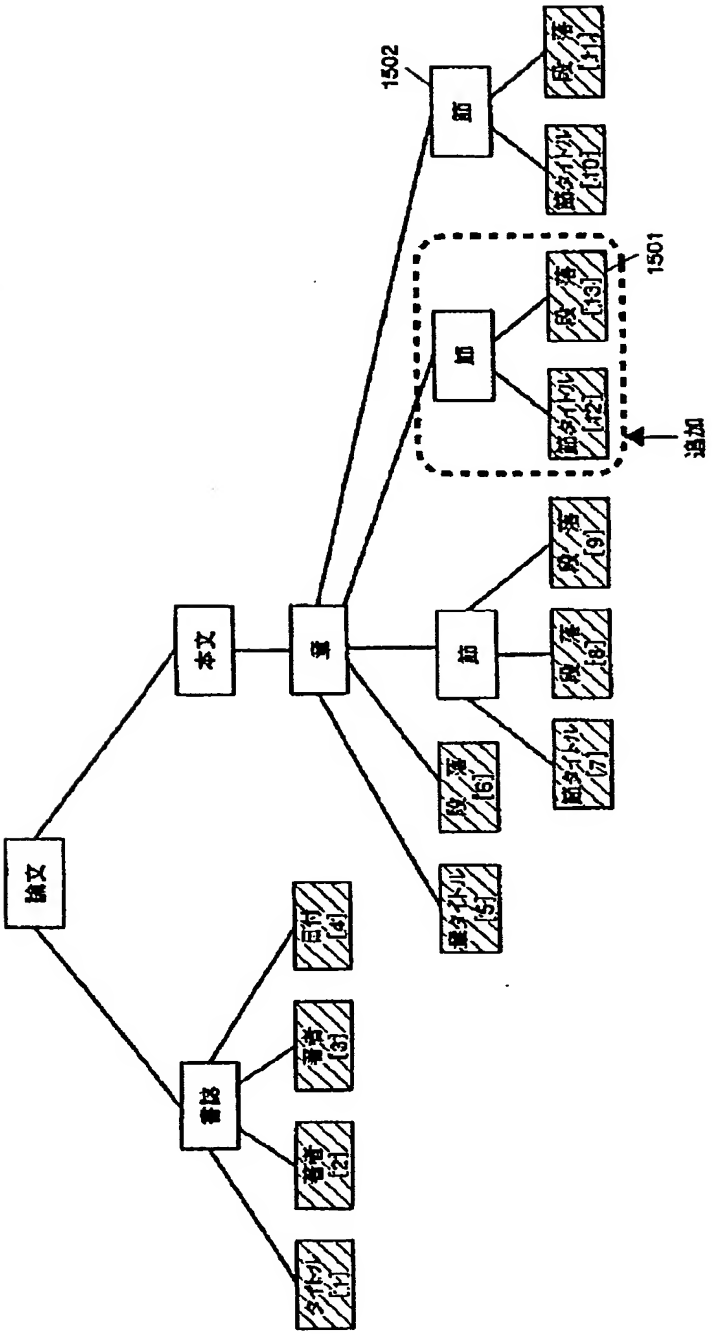
【図３５】


文書識別子
文脈識別子
内容文字列
D1
C129
"変換処理の実例"
2904
部分文字列
2905
文書識別子
2906
文脈識別子
2907
文字位置
"変換"
"換処"
"処理"
"理の"
"の実"
"実例"
D1
C129
X+1
X+2
X+3
X+4
X+5
X+6
文字列インデックス

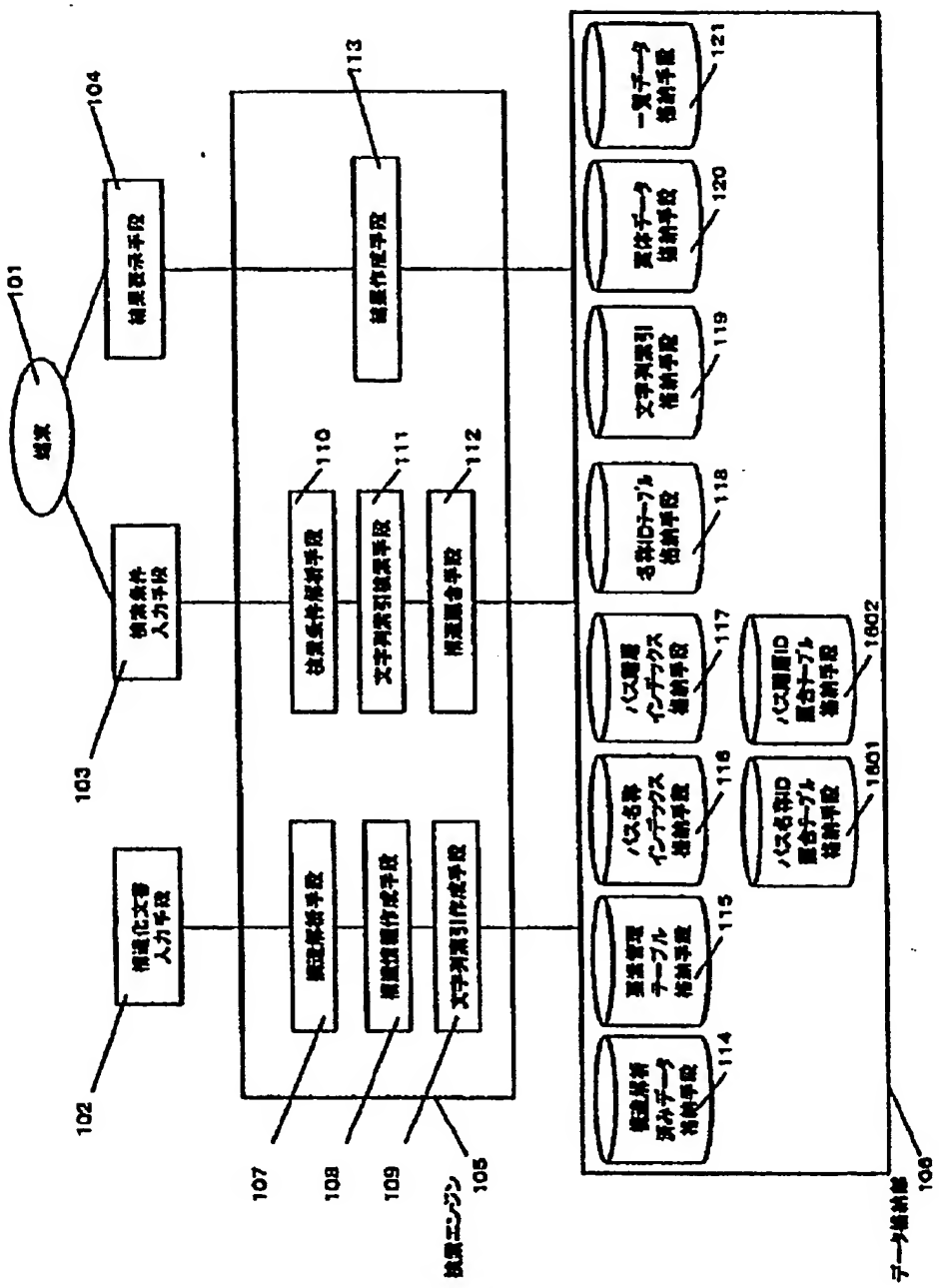
【図１４】


構造指定
(パス名称)
/論文/書誌/タイトル
パス名称ID
N2
要素管理テーブル
検索単位
識別子
文書番号
パス名称ID
パス階層ID
名称ID
1
1
N2
L2
T3
:
...
...
...
...
8
1
N11
L4
T9
:
...
...
...
...
文字列索引検索結果群
(検索単位識別子)
1
8
1401
検索結果
検索単位
識別子
文書番号
1
1

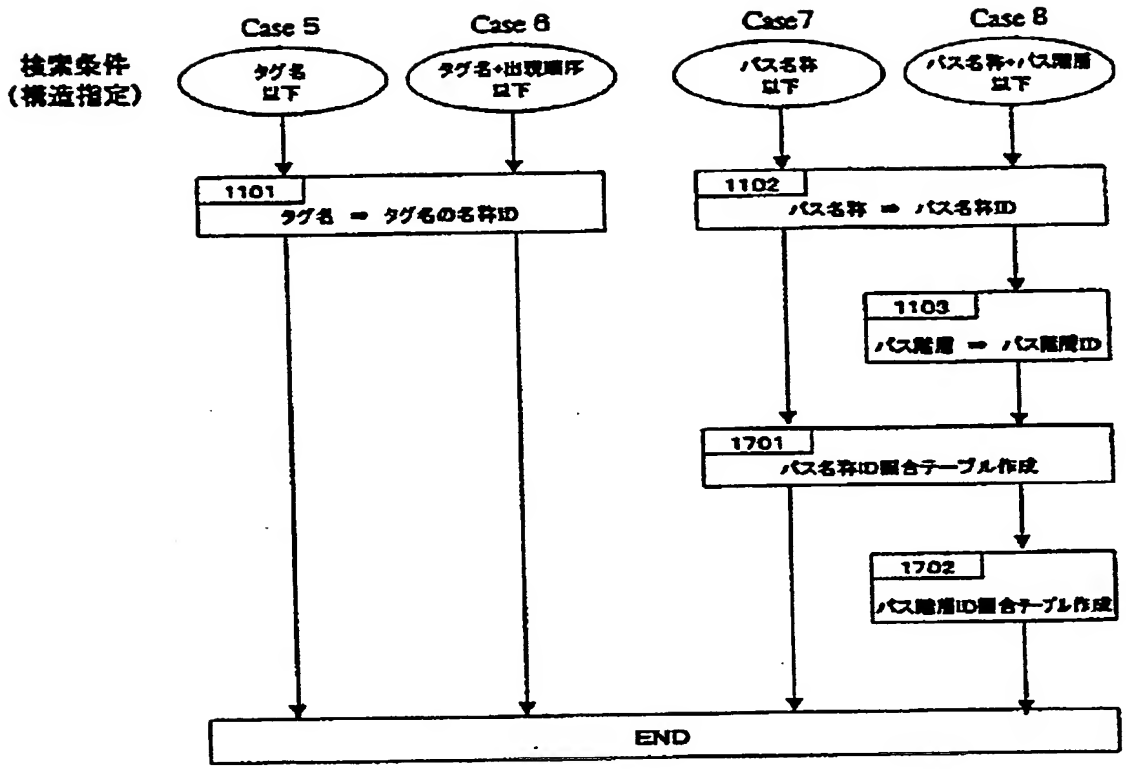
【図１５】


論文
書誌
目次 [4]
著者 [3]
書誌 [2]
タイトル [1]
本文
章
段落 [5]
節
段落 [6]
節タイトル [7]
節
段落 [9]
段落 [8]
節タイトル [10]
1502
節
段落 [11]
節タイトル [101]
1501
節
段落 [13]
節タイトル [12]
追加

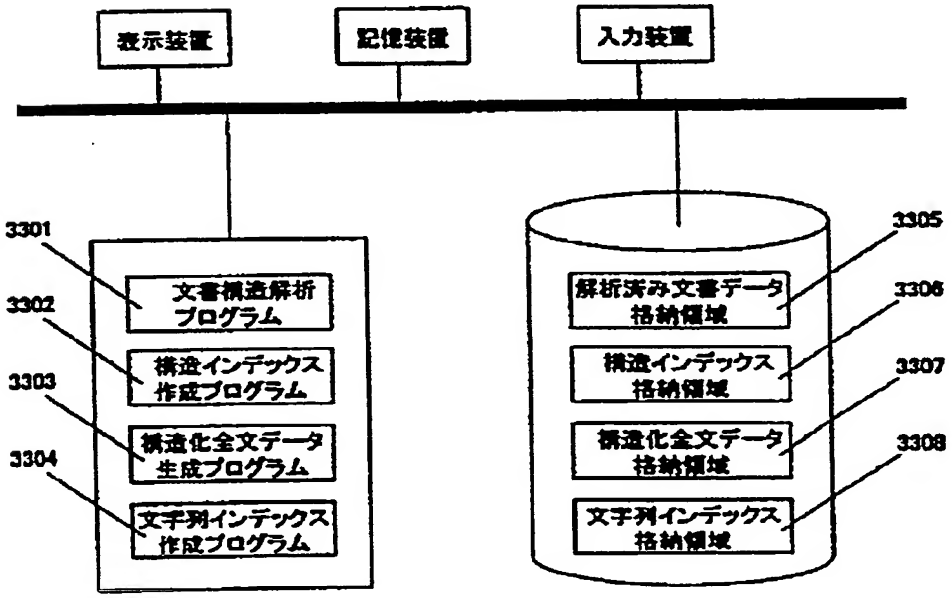
【図１６】


101
結果
104
結果表示手段
103
検索条件入力手段
102
概念化文書入力手段
113
結果作成手段
110
検索条件展開手段
111
文字列索引検索手段
112
概念照合手段
107
概念展開手段
108
概念索引作成手段
109
文字列索引作成手段
検索エンジン
105
121
一覧データ格納手段
120
実体データ格納手段
119
文字列索引格納手段
118
名称IDテーブル格納手段
117
バス属性インデックス格納手段
1002
バス属性ID照合テーブル格納手段
116
バス名称インデックス格納手段
1001
バス名称ID照合テーブル格納手段
115
概念管理テーブル格納手段
114
概念展開済みデータ格納手段
106
データ格納部

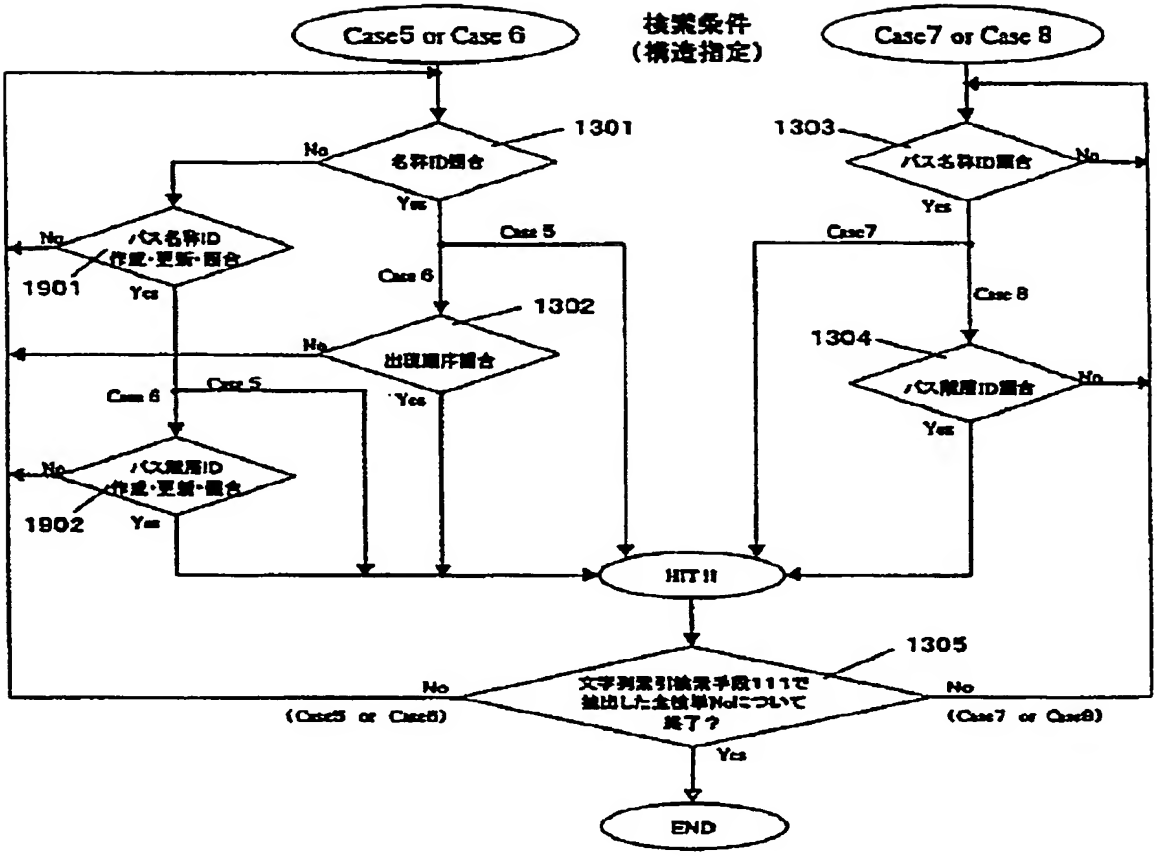
【図１７】


検索条件
(構造指定)
Case 5
タグ名
以下
Case 6
タグ名+出現順序
以下
Case7
パス名称
以下
Case 8
パス名称+パス階層
以下
1101
タグ名 ⇒ タグ名のID
1102
パス名称 ⇒ パスID
1103
パス階層 ⇒ パス階層ID
1701
パスID組合テーブル作成
1702
パス階層ID組合テーブル作成
END

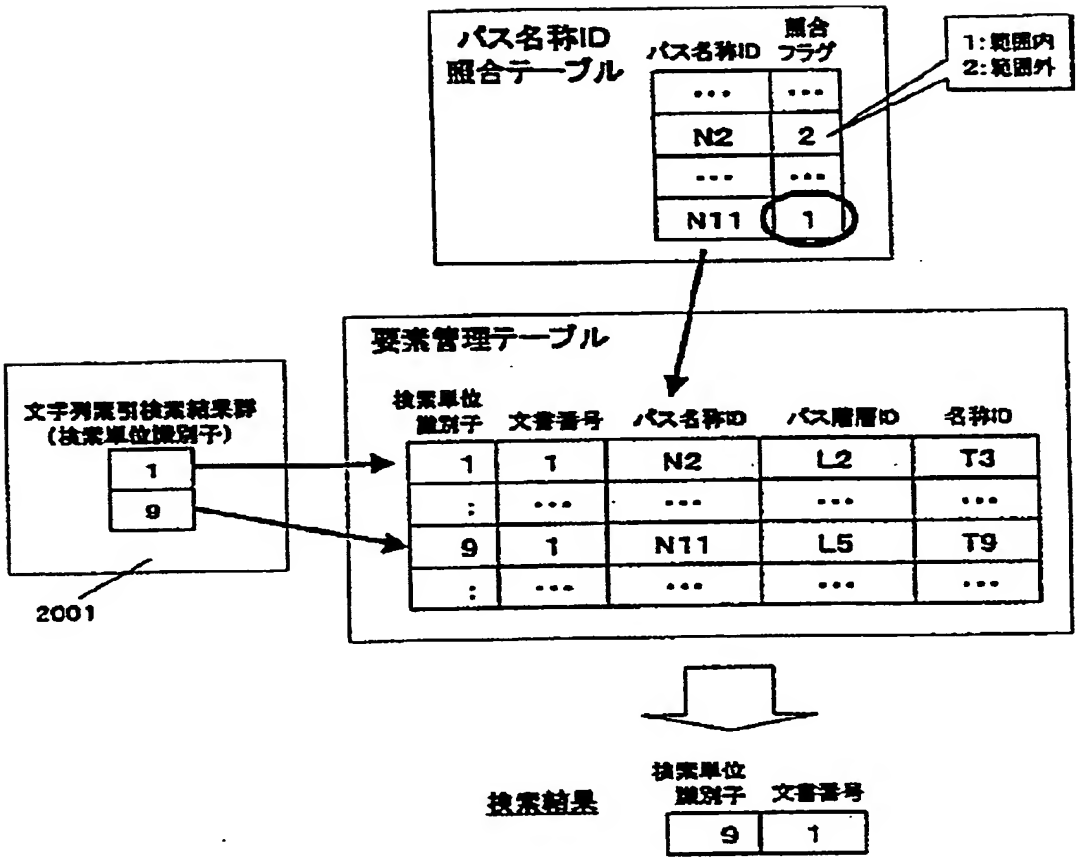
【図３３】


表示装置
記憶装置
入力装置
3301
文書構造解析
プログラム
3302
構造インデックス
作成プログラム
3303
構造化全文データ
生成プログラム
3304
文字列インデックス
作成プログラム
3305
解析済み文書データ
格納領域
3306
構造インデックス
格納領域
3307
構造化全文データ
格納領域
3308
文字列インデックス
格納領域

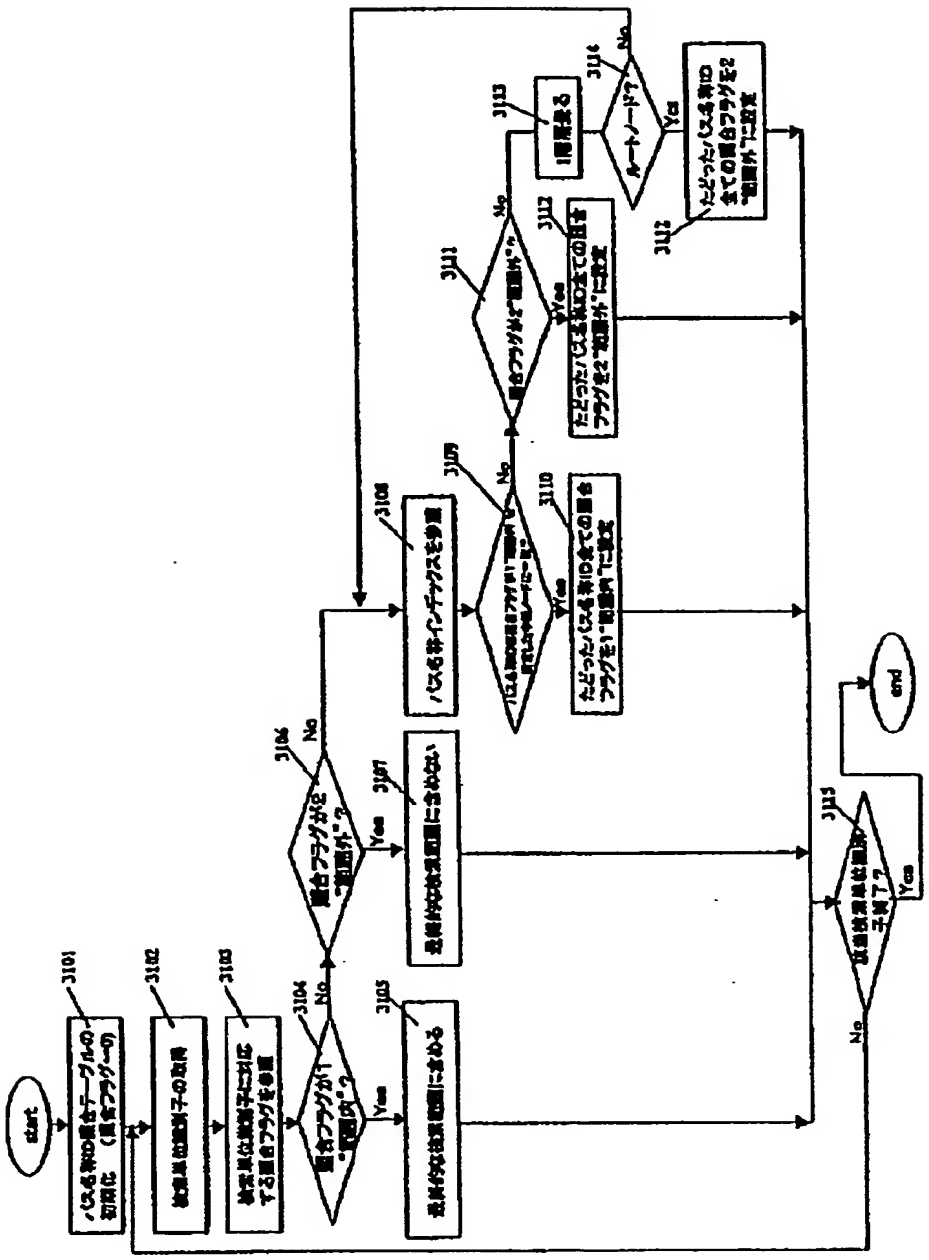
【図１９】


Case5 or Case 6
検索条件
(構造指定)
Case7 or Case 8
1301
名称ID照合
No
Yes
1303
パス名称ID照合
No
Yes
パス名称ID
作成・更新・照合
No
Yes
1901
Case 5
Case 6
Case7
Case 8
1302
出現順序照合
No
Yes
1304
パス階層ID照合
No
Yes
Case 6
Case 5
パス階層ID
作成・更新・照合
No
Yes
1902
HIT !!
1305
文字列索引検索手段111で
抽出した全検索Noについて
終了？
No
(Case5 or Case6)
No
(Case7 or Case8)
Yes
END

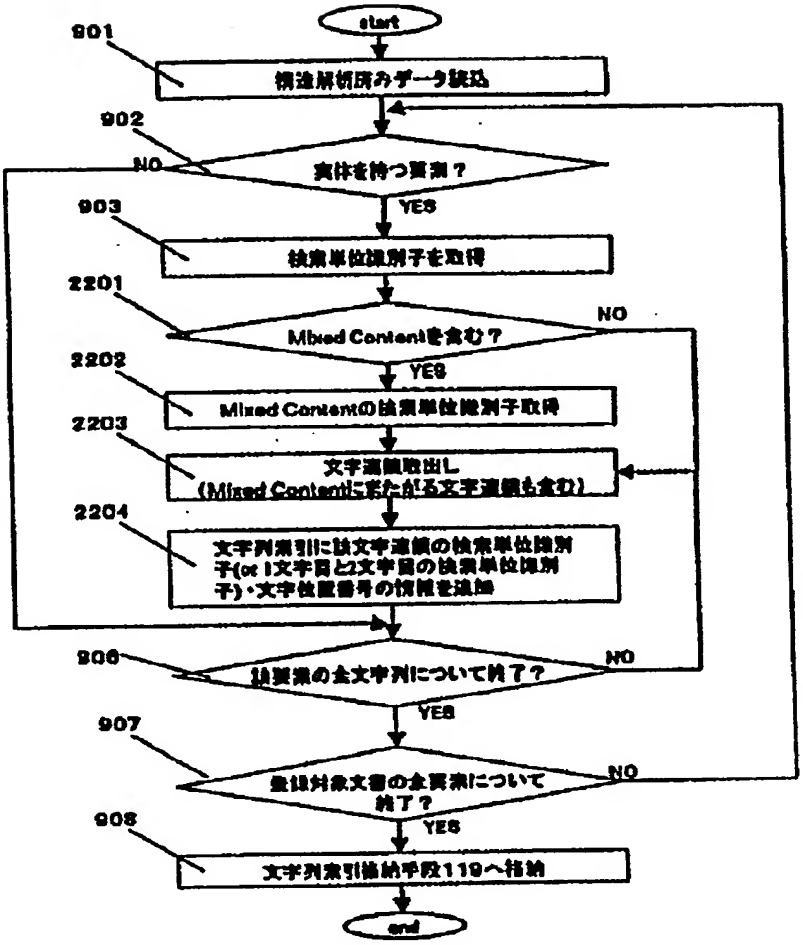
【図２０】


パス名称ID
照合テーブル
パス名称ID
照合
フラグ
N2
2
N11
1
1:範囲内
2:範囲外
要素管理テーブル
検索単位
識別子
文書番号
パス名称ID
パス階層ID
名称ID
1
1
N2
L2
T3
9
1
N11
L5
T9
文字列索引検索結果群
(検索単位識別子)
1
9
2001
検索結果
検索単位
識別子
文書番号
9
1

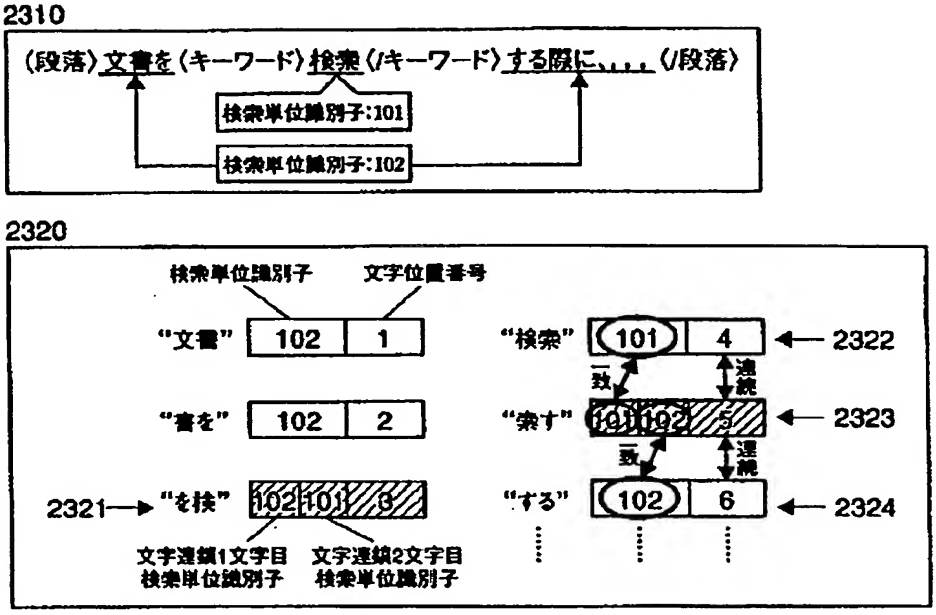
【図２１】


start
3101
3102
3103
3104
No
Yes
3105
3106
No
Yes
3107
3108
3109
No
Yes
3110
3111
No
Yes
3112
3113
3114
No
Yes
3115
No
Yes
end

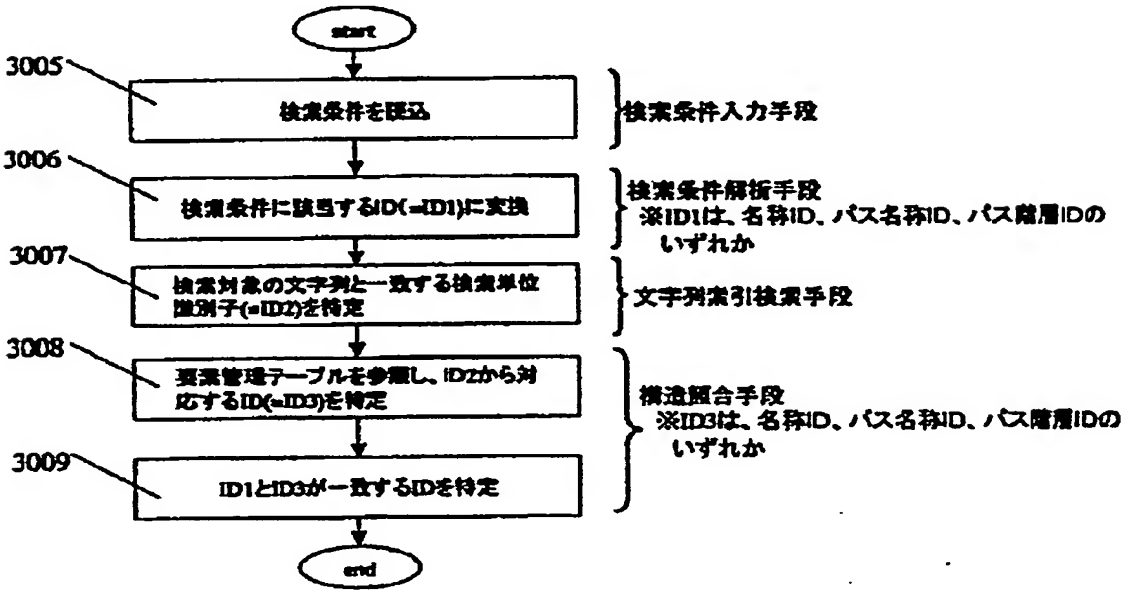
【図２３】


start
901
構造解析済みデータ読込
902
実体を持つ要素？
NO
YES
903
検索単位識別子を取得
2201
Mixed Contentを含む？
NO
YES
2202
Mixed Contentの検索単位識別子取得
2203
文字連鎖取出し
(Mixed Contentにまたがる文字連鎖も含む)
2204
文字列索引に該文字連鎖の検索単位識別子(or 1文字目と2文字目の検索単位識別子)・文字位置番号の情報を追加
906
該要素の全文字列について終了？
NO
YES
907
登録対象文書の全要素について終了？
NO
YES
908
文字列索引格納手段119へ格納
end

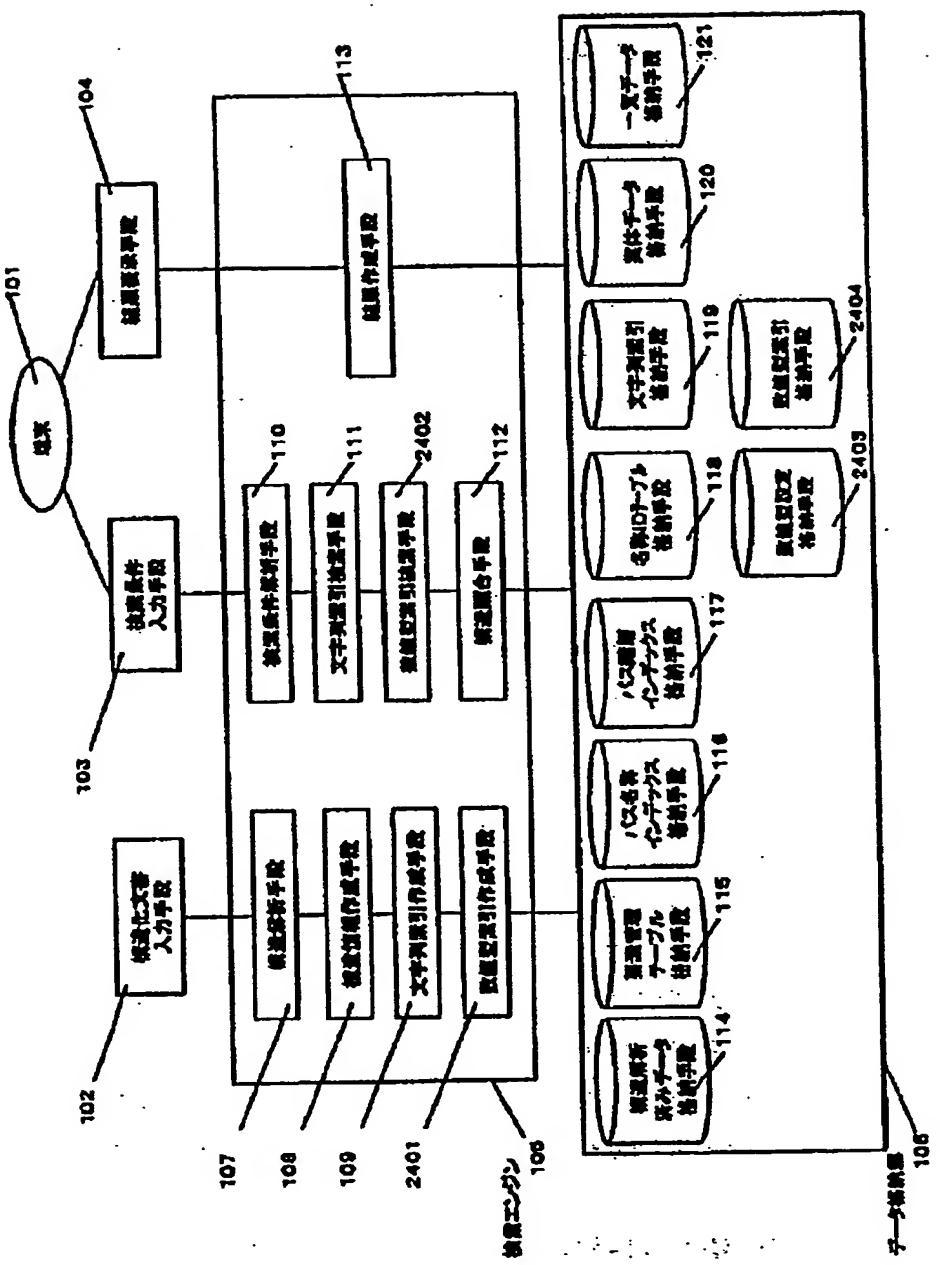
【図２４】


2310
〈段落〉文書を〈キーワード〉検索〈/キーワード〉する際に、、、、〈/段落〉
検索単位識別子:101
検索単位識別子:102
2320
検索単位識別子
文字位置番号
"文書"
102
1
"書を"
102
2
2321
"を検"
102
101
3
文字連鎖1文字目 検索単位識別子
文字連鎖2文字目 検索単位識別子
"検索"
101
4
2322
一致
連続
"索す"
101
102
5
2323
一致
連続
"する"
102
6
2324

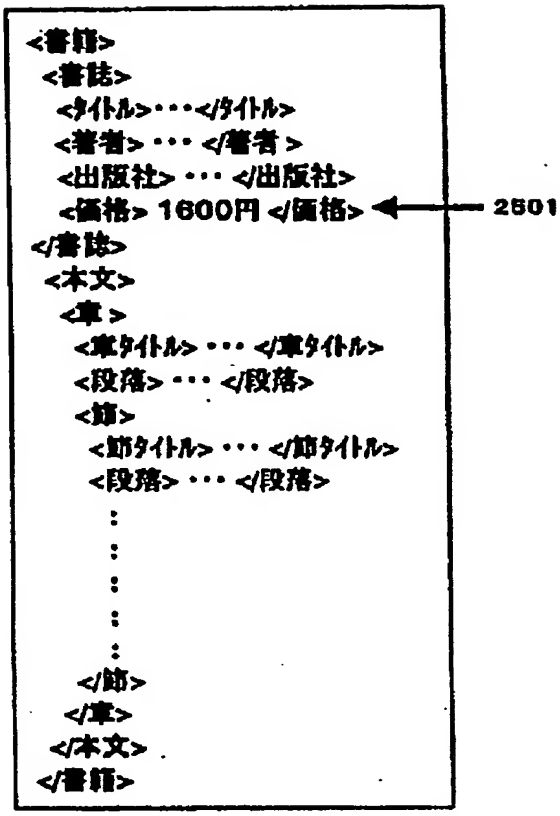
【図３０】


start
3005
検索条件を取込
検索条件入力手段
3006
検索条件に該当するID(=ID1)に変換
検索条件解析手段
※ID1は、名称ID、パス名称ID、パス階層IDのいずれか
3007
検索対象の文字列と一致する検索単位識別子(=ID2)を特定
文字列索引検索手段
3008
要素管理テーブルを参照し、ID2から対応するID(=ID3)を特定
3009
ID1とID3が一致するIDを特定
構造照合手段
※ID3は、名称ID、パス名称ID、パス階層IDのいずれか
end

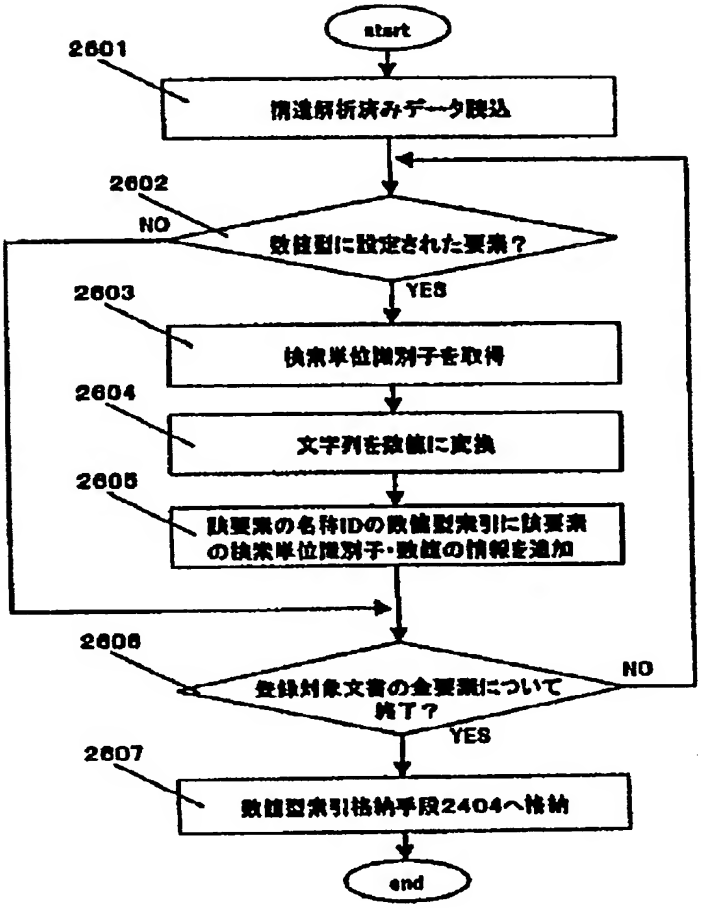
【図２５】


101
104
113
121
120
119
2404
110
111
2402
112
2403
118
117
116
115
114
103
102
107
108
109
2401
106
105

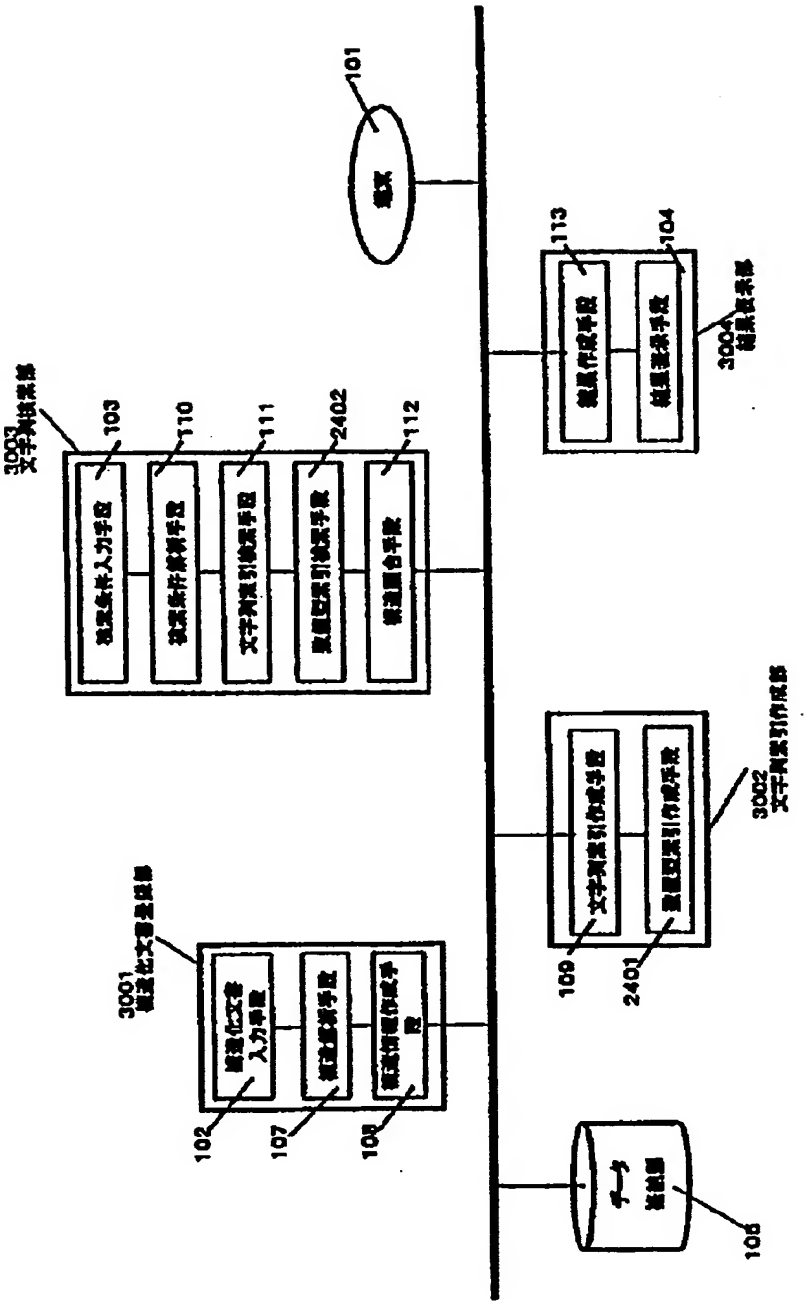
【図２６】


<書籍>
<書誌>
<タイトル>・・・</タイトル>
<著者>・・・</著者>
<出版社>・・・</出版社>
<価格> 1600円 </価格>
2501
</書誌>
<本文>
<章>
<章タイトル>・・・</章タイトル>
<段落>・・・</段落>
<節>
<節タイトル>・・・</節タイトル>
<段落>・・・</段落>
</節>
</章>
</本文>
</書籍>

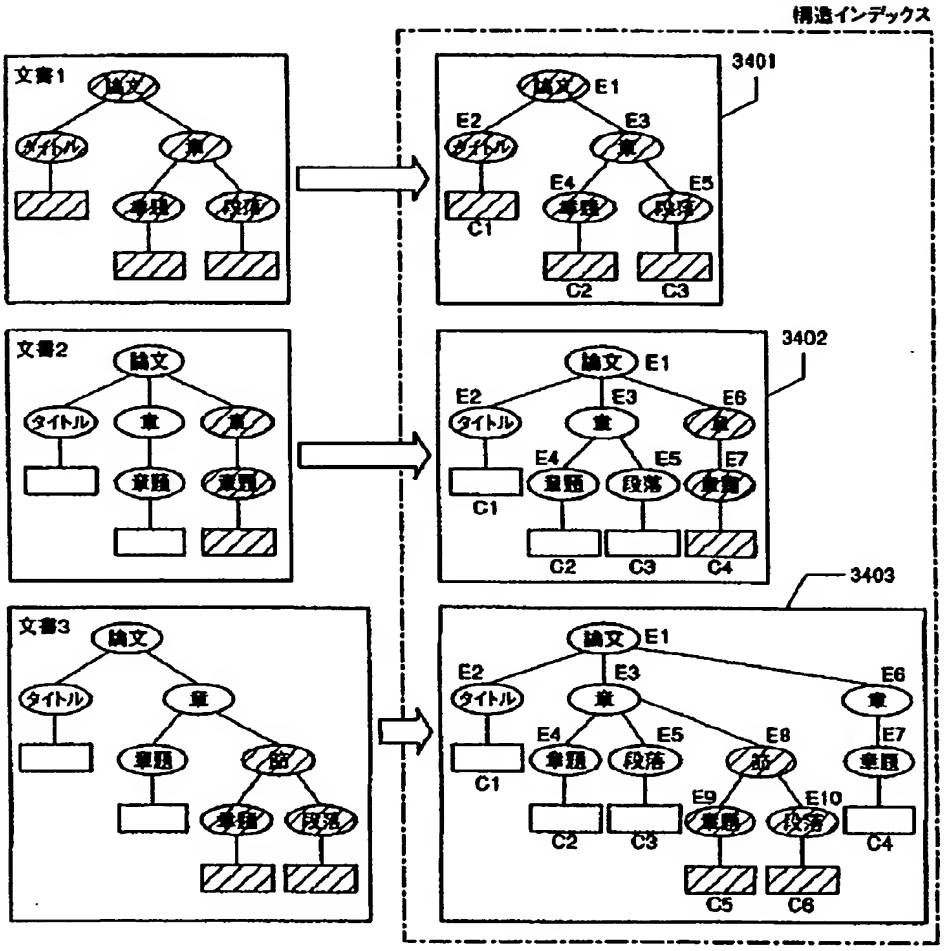
【図２７】


start
2601
構造解析済みデータ読込
2602
NO
数値型に設定された要素？
YES
2603
検索単位識別子を取得
2604
文字列を数値に変換
2605
該要素の名称IDの数値型索引に該要素
の検索単位識別子・数値の情報を追加
2606
NO
登録対象文書の全要素について
終了？
YES
2607
数値型索引格納手段2404へ格納
end

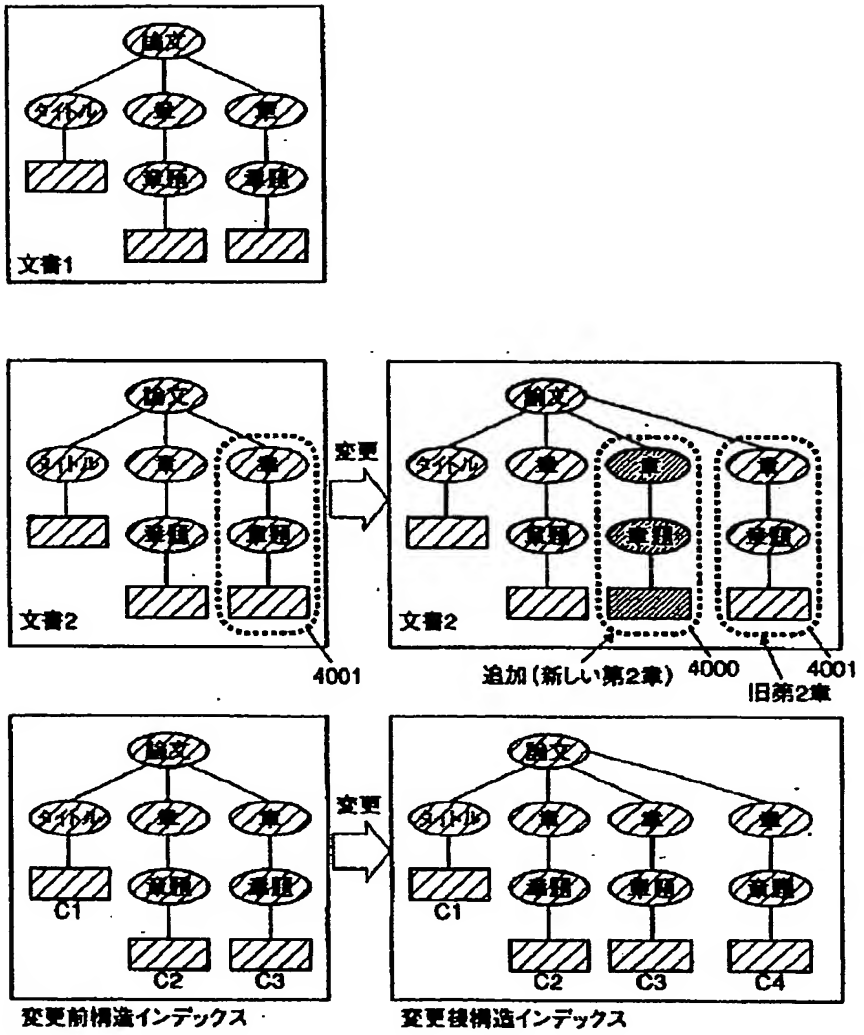
【図２９】



【図３４】


構造インデックス
文書1
論文
タイトル
章
章題
段落
3401
E1
E2
E3
E4
E5
C1
C2
C3
文書2
3402
E6
E7
C4
文書3
3403
E8
E9
E10
C5
C6

【図３６】



フロントページの続き

(72)発明者　鶴林　健
大阪府門真市大字門真1006番地　松下電器産業株式会社内

(72)発明者　片山　修
大阪府門真市大字門真1006番地　松下電器産業株式会社内

(72)発明者　中井　信一
大阪府門真市大字門真1006番地　松下電器産業株式会社内

Fターム(参考)　5B075 ND35 NK43